FORESTRY RESEARCH SUPPORT PROGRAMME

FOR ASIA AND THE PACIFIC

A Statistical Manual For Forestry Research

FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS
REGIONAL OFFICE FOR ASIA AND THE PACIFIC
BANGKOK

May 1999

FORESTRY RESEARCH SUPPORT PROGRAMME

FOR ASIA AND THE PACIFIC

A STATISTICAL MANUAL FOR FORESTRY RESEARCH

By

K. JAYARAMAN
Kerala Forest Research Institute
Peechi, Thrissur, Kerala, India

FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS
REGIONAL OFFICE FOR ASIA AND THE PACIFIC
BANGKOK

ACKNOWLEDGEMENTS

The author is deeply indebted to the FORSPA for having supported the preparation of this manual. The author is also indebted to the Kerala Forest Research Institute for having granted permission to undertake this work and offering the necessary infrastructure facilities. Many examples used for illustration of the different statistical techniques described in the manual were based on data generated by Scientists at the Kerala Forest Research Institute. The author extends his gratitude to all his colleagues at the Institute for having gracefully co-operated in this regard. The author also wishes to thank deeply Smt. C. Sunanda and Mr. A.G. Varghese, Research Fellows of the Division of Statistics, Kerala Forest Research Institute, for patiently going through the manuscript and offering many helpful suggestions to improve the same in all respects.

This manual is dedicated to those who are determined to seek TRUTH, cutting down the veil of chance by the sword of pure reason.

March, 1999 *K. Jayaraman*

INTRODUCTION

This manual was written on a specific request from FORSPA, Bangkok to prepare a customised training manual supposed to be useful to researchers engaged in forestry research in Bhutan. To that effect, a visit was made to Bhutan to review the nature of the research investigations being undertaken there and an outline for the manual was prepared in close discussion with the researchers. Although the content of the manual was originally requested to be organised in line with the series of research investigations planned under the Eighth Five Year Plan for Bhutan, the format was so designed that the manual is useful to a wider set of researchers engaged in similar investigations. The manual is intended to be a source of reference for researchers engaged in research on renewable natural resources especially forests, agricultural lands and livestock, in designing research investigations, collecting and analysing relevant data and also in interpreting the results. The examples used for illustration of various techniques are mainly from the field of forestry.

After some introductory remarks on the nature of scientific method and the role of statistics in scientific research, the manual deals with specific statistical techniques starting from basic statistical estimation and testing procedures, methods of designing and analysing experiments and also some standard sampling techniques. Further, statistical methods involved in certain specific fields like tree breeding, wildlife biology, forest mensuration and ecology many of which are unique to forestry research, are described.

The description of the methods is not exhaustive because there is always a possibility of utilizing the data further depending on the needs of the investigators and also because refinements in methodology are happening continuously. The intention of the manual has been more on introducing the researchers to some of the basic concepts and techniques in statistics which have found wide application in research in forestry and allied fields.

There was also a specification that the manual is to be written in as simple a manner as possible with illustrations so that it serves as a convenient and reference manual for the actual researchers. For these reasons, description of only simple modes of design and analysis are given with appropriate illustrations. More complicated techniques available are referred to standard text books dealing with the topics. However, every effort has been made to include in the manual as much of material required for a basic course in applied statistics indicating several areas of application and directions for further reading. Inclusion of additional topics would have made it just unwieldy.

Any body with an elementary knowledge in basic mathematics should be able to follow successfully the description provided in the manual. To the extent possible, calculus and matrix theory are avoided and where unavoidable, necessary explanation is offered. For a beginner, the suggested sequence for reading is the one followed for the different chapters in the manual. More experienced researchers can just skim through the initial sections and start working on the applications discussed in the later sections.

NOTATION

Throughout this book, names of variables are referred in italics. The symbol $\sum$ is used to represent 'the sum of'. For example, the expression $G = y_1 + y_2 + \ldots + y_n$ can be written as $G = \sum_{i=1}^{n} y_i$ or simply $G = \sum y$ when the range of summation is understood from the context.

In the case of summation involving multiple subscripts, the marginal sums are denoted by placing a dot (.) over that subscript, like,

$$\sum_j y_{ij} = y_{i.}, \quad \sum_i y_{ij} = y_{.j}, \sum_{ij} y_{ij} = y_{..}$$

When two letters are written side by side such as *ab*, in equations, it generally indicates product of *a* and *b* unless otherwise specified or understood from the context. Multiplication with numerals are indicated by brackets *e.g.*, (4)(5) would mean 4 multiplied 5. Division is indicated either by slash (/) or a horizontal mid-line between the numerator and the denominator.

Equations, tables and figures are numbered with reference to the chapter numbers. For instance, Equation (3.1) refers to equation 1 in chapter 3.

Certain additional notations such as those pertaining to factorial notation, combinatorics, matrices and related definitions are furnished in Appendix 7.

1. STATISTICAL METHOD IN SCIENTIFIC RESEARCH

Like in any other branch of science, forestry research is also based on scientific method which is popularly known as the inductive-deductive approach. Scientific method entails formulation of hypotheses from observed facts followed by deductions and verification repeated in a cyclical process. Facts are observations which are taken to be true. Hypothesis is a tentative conjecture regarding the phenomenon under consideration. Deductions are made out of the hypotheses through logical arguments which in turn are verified through objective methods. The process of verification may lead to further hypotheses, deductions and verification in a long chain in the course of which scientific theories, principles and laws emerge.

As a case of illustration, one may observe that trees in the borders of a plantation are growing better than trees inside. A tentative hypothesis that could be formed from this fact is that the better growth of trees in the periphery is due to increased availability of light from the open sides. One may then deduce that by varying the spacing between trees and thereby controlling the availability of light, the trees can be made to grow differently. This would lead to a spacing experiment wherein trees are planted at different espacements and the growth is observed. One may then observe that trees under the same espacement vary in their growth and a second hypothesis formed would be that the variation in soil fertility is the causative factor for the same. Accordingly, a spacing cum fertilizer trial may follow. Further observation that trees under the same espacement, receiving the same fertilizer dose differ in their growth may prompt the researcher to conduct a spacing cum fertilizer cum varietal trial. At the end of a series of experiments, one may realize that the law of limiting factors operate in such cases which states that crop growth is constrained by the most limiting factor in the environment.

The two main features of scientific method are its repeatability and objectivity. Although this is rigorously achieved in the case of many physical processes, biological phenomena are characterised by variation and uncertainty. Experiments when repeated under similar conditions need not yield identical results, being subjected to fluctuations of random nature. Also, observations on the complete set of individuals in the population are out of question many times and inference may have to be made quite often from a sample set of observations. The science of statistics is helpful in objectively selecting a sample, in making valid generalisations out of the sample set of observations and also in quantifying the degree of uncertainty in the conclusions made.

Two major practical aspects of scientific investigations are collection of data and interpretation of the collected data. The data may be generated through a sample survey on a naturally existing population or a designed experiment on a hypothetical population. The collected data are condensed and useful information extracted through techniques of statistical inference. This apart, a method of considerable importance to forestry which has gained wider acceptance in recent times with the advent of computers is simulation. This is particularly useful in forestry because simulation techniques can replace large scale field experiments which are extremely costly and time consuming. Mathematical models are developed which capture most of the

relevant features of the system under consideration after which experiments are conducted in computer rather than with real life systems. A few additional features of these three approaches *viz*., survey, experiment and simulation are discussed here before describing the details of the techniques involved in later chapters.

In a broad sense, all *in situ* studies involving non-interfering observations on nature can be classed as surveys. These may be undertaken for a variety of reasons like estimation of population parameters, comparison of different populations, study of the distribution pattern of organisms or for finding out the interrelations among several variables. Observed relationships from such studies are not many times causative but will have predictive value. Studies in sciences like economics, ecology and wildlife biology generally belong to this category. Statistical theory of surveys relies on random sampling which assigns known probability of selection for each sampling unit in the population.

Experiments serve to test hypotheses under controlled conditions. Experiments in forestry are held in forests, nurseries and laboratories with pre-identified treatments on well defined experimental units. The basic principles of experimentation are randomization, replication and local control which are the prerequisites for obtaining a valid estimate of error and for reducing its magnitude. Random allocation of the experimental units to the different treatments ensures objectivity, replication of the observations increases the reliability of the conclusions and the principle of local control reduces the effect of extraneous factors on the treatment comparison. Silvicultural trials in plantations and nurseries and laboratory trials are typical examples of experiments in forestry.

Experimenting on the state of a system with a model over time is termed simulation. A system can be formally defined as a set of elements also called components. A set of trees in a forest stand, producers and consumers in an economic system are examples of components. The elements (components) have certain characteristics or attributes and these attributes have numerical or logical values. Among the elements, relationships exist and the consequently, the elements are interacting. The state of a system is determined by the numerical or logical values of the attributes of the system elements. The interrelations among the elements of a system are expressible through mathematical equations and thus the state of the system under alternative conditions is predictable through mathematical models. Simulation amounts to tracing the time path of a system under alternative conditions.

While surveys and experiments and simulations are essential elements of any scientific research programme, they need to be embedded in some larger and more strategic framework if the programme as a whole is to be both efficient and effective. Increasingly, it has come to be recognized that systems analysis provides such a framework, designed to help decision makers to choose a desirable course of action or to predict the outcome of one or more courses of action that seems desirable. A more formal definition of systems analysis is the orderly and logical organisation of data and information into models followed by rigorous testing and exploration of these models necessary for their validation and improvement (Jeffers ,1978).

Research related to forests extends from molecular level to the whole of biosphere. The nature of the material dealt with largely determines the methods employed for making investigations. Many levels of organization in the natural hierarchy such as micro-organisms or trees are amenable to experimentation but only passive observations and modelling are possible at certain other levels. Regardless of the objects dealt with, the logical framework of the scientific approach and the statistical inference can be seen to remain the same. This manual essentially deals with various statistical methods used for objectively collecting the data and making valid inferences out of the same.

2. BASIC STATISTICS

2.1. Concept of probability

The concept of probability is central to the science of statistics. As a subjective notion, probability can be interpreted as degree of belief in a continuous range between impossibility and certainty, about the occurrence of an event. Roughly speaking, the value *p*, given by a person for the probability P(E) of an event E, means the price that person is willing to pay for winning a fixed amount of money conditional on the event being materialized. If the price the person is willing to pay is x units for winning y units of money, then the probability assigned is indicated by P(E)= x / (x + y). More objective measures of probability are based on equally likely outcomes and that based on relative frequency which are described below. A rigorous axiomatic definition of probability is also available in statistical theory which is not dealt with here.

Classical definition of probability : Suppose an event E can happen in x ways out of a total of n possible equally likely ways. Then the probability of occurrence of the event (called its success) is denoted by

$$p = \mathrm{P(E)} = \frac{x}{n} \tag{2.1}$$

The probability of non-occurrence of the event (called its failure) is denoted by

$$q = \mathrm{P(not\ E)} = \frac{n-x}{n} = 1-\frac{x}{n} \tag{2.2}$$

$$= 1-p = 1-\mathrm{P(E)} \tag{2.3}$$

Thus $p + q = 1$, or P(E) + P(not E) = 1. The event 'not E' is sometimes denoted by $\overline{\mathrm{E}}$, $\tilde{\mathrm{E}}$ or ~ E .

As an example, let the colour of flowers in a particular plant species be governed by the presence of a dominant gene A in a single gene locus, the gametic combinations AA and Aa giving rise to red flowers and the combination aa giving white flowers. Let E be the event of getting red flowers in the progeny obtained through selfing of a heterozygote, Aa. Let us assume that the four gametic combinations AA, Aa, aA and aa are equally likely. Since the event E can occur in three of these ways, we have,

$$p = P(E) = \frac{3}{4}$$

The probability of getting white flowers in the progeny through selfing of the heterozygote Aa is

$$q = P(\overline{E}) = 1 - \frac{3}{4} = \frac{1}{4}$$

Note that the probability of an event is a number between 0 and 1. If the event cannot occur, its probability is 0. If it must occur, *i.e.*, its occurrence is certain, its probability is 1. If p is the probability that an event will occur, the odds in favour of its happening are $p:q$ (read 'p to q'); the odds against its happening are $q:p$. Thus the odds in favour of red flowers in the above example are

$$p:q = \frac{3}{4}:\frac{1}{4} = 3:1, \text{ } i.e. \text{ 3 to 1.}$$

Frequency interpretation of probability : The above definition of probability has a disadvantage in that the words 'equally likely' are vague. Since these words seem to be synonymous with 'equally probable', the definition is circular because, we are essentially defining probability in terms of itself. For this reason, a statistical definition of probability has been advocated by some people. According to this, the estimated probability, or empirical probability, of an event is taken as the relative frequency of occurrence of the event when the number of observations is large. The probability itself is the limit of the relative frequency as the number of observations increases indefinitely. Symbolically, probability of event E is,

$$P(E) = \lim_{n \to \infty} f_n(E) \qquad (2.4)$$

where $f_n(E)$ = (number of times E occurred)/(total number of observations)

For example, in a search for a particular endangered species, the following numbers of plants of that species were encountered in a survey in sequence.

x (number of plants of endangered species) :	1,	6,	62,	610
n (number of plants examined) :	1000,	10000,	100000,	1000000
p (proportion of endangered species) :	0.001,	0.00060,	0.00062,	0.00061

As n tends to infinity, the relative frequency seems to approach a certain limit. We call this empirical property as the stability of the relative frequency.

Conditional probability, independent and dependent events : If E_1 and E_2 are two events, the probability that E_2 occurs given that E_1 has occurred is denoted by $P(E_2/E_1)$ or $P(E_2$ given $E_1)$ and is called the conditional probability of E_2 given that E_1 has occurred. If the occurrence or non-occurrence of E_1 does not affect the probability of occurrence of E_2 then $P(E_2/E_1) = P(E_2)$ and we say that E_1 and E_2 are independent events; otherwise they are dependent events.

If we denote by E_1E_2 the event that 'both E_1 and E_2 occur', sometimes called a compound event, then

$$P(E_1E_2) = P(E_1)P(E_2/E_1) \quad (2.5)$$

In particular, $P(E_1E_2) = P(E_1)P(E_2)$ for independent events. (2.6)

For example, consider the joint segregation of two characters *viz*., flower colour and shape of seed in a plant species, the characters being individually governed by the presence of dominant genes A and B respectively. Individually, the combinations AA and Aa give rise to red flowers and the combination aa give white flowers, the combinations BB and Bb give round seeds and the combination bb produce wrinkled seeds.

Let E_1 and E_2 be the events of 'getting plants with red flowers' and 'getting plants with round seeds' in the progeny obtained through selfing of a heterozygote AaBb respectively. If E_1 and E_2 are independent events, *i.e*., there is no interaction between the two gene loci, the probability of getting plants with red flowers and round seeds in the selfed progeny is,

$$P(E_1E_2)=P(E_1)P(E_2)=\left(\frac{3}{4}\right)\left(\frac{3}{4}\right) = \frac{9}{16}$$

In general, if E_1, E_2, E_3, …, E_n are n independent events having respective probabilities p_1, p_2, p_3, …, p_n, then the probability of occurrence of E_1 and E_2 and E_3 and … E_n is $p_1p_2p_3\ldots p_n$.

2.2. Frequency distribution

Since the frequency interpretation of probability is highly useful in practice, preparation of frequency distribution is an often-used technique in statistical works when summarising large masses of raw data, which leads to information on the pattern of occurrence of predefined classes of events. The raw data consist of measurements of some attribute on a collection of individuals. The measurement would have been made in one of the following scales *viz*., nominal, ordinal, interval or ratio scale. Nominal scale refers to measurement at its weakest level when number or other symbols are used simply to classify an object, person or characteristic, *e.g.*, state of health (healthy, diseased). Ordinal scale is one wherein given a group of equivalence classes, the relation greater than holds for all pairs of classes so that a complete rank ordering of classes is possible, *e.g*., socio-economic status. When a scale has all the characteristics of an ordinal scale, and when in addition, the distances between any two numbers on the scale are of known size, interval scale is achieved,. e*g*., temperature scales like centigrade or Fahrenheit. An interval scale with a true zero point as its origin forms a ratio scale. In a ratio scale, the ratio of any two scale points is independent of the unit of measurement, *e.g*., height of trees. Reference may be made to Siegel (1956) for a

detailed discussion on the different scales of measurement, their properties and admissible operations in each scale.

Regardless of the scale of measurement, a way to summarise data is to distribute it into *classes* or *categories* and to determine the number of individuals belonging to each class, called the *class frequency.* A tabular arrangement of data by classes together with the corresponding class frequencies is called a *frequency distribution* or *frequency table*. Table 2.1 is a frequency distribution of diameter at breast-height (dbh) recorded to the nearest cm, of 80 teak trees in a sample plot. The *relative frequency* of a class is the frequency of the class divided by the total frequency of all classes and is generally expressed as a percentage. For example, the relative frequency of the class 17-19 in Table 2.1 is (30/80)100 = 37.4%. The sum of all the relative frequencies of all classes is clearly 100 %.

Table 2.1. Frequency distribution of dbh of teak trees in a plot

Dbh class (cm)	Frequency (Number of trees)	Relative frequency (%)
11-13	11	13.8
14-16	20	25.0
17-19	30	37.4
20-22	15	18.8
23-25	4	5.0
Total	80	100.0

A symbol defining a class interval such as 11-13 in the above table is called a *class interval*. The end numbers 11 and 13, are called *class limits*; the smaller number 11 is the *lower class limit* and the larger number 13 is the *upper class limit*. The terms class and class interval are often used interchangeably, although the class interval is actually a symbol for the class. A class interval which, at least theoretically, has either no upper class limit or no lower class limit indicated is called an *open class interval*. For example, the class interval '23 cm and over' is an open class interval.

If dbh values are recorded to the nearest cm, the class interval 11-13, theoretically includes all measurements from 10.5 to 13.5 cm. These numbers are called class boundaries or true class limits; the smaller number 10.5 is the *lower class boundary* and the large number 13.5 is the *upper class boundary*. In practice, the class boundaries are obtained by adding the upper limit of one class interval to the lower limit of the next higher class interval and dividing by 2.

Sometimes, class boundaries are used to symbolise classes. For example, the various classes in the first column of Table 2.1 could be indicated by 10.5-13.5, 13.5-16.5, etc. To avoid ambiguity in using such notation, class boundaries should not coincide with actual observations. Thus, if an observation were 13.5 it would not be possible to decide whether it belonged to the class interval 10.5-13.5 or 13.5-16.5. The size or width of a class interval is the difference between the lower and upper boundaries and is also referred as the *class width.* The class mark is the midpoint of the class interval and is obtained by adding the lower and upper class limits and dividing by two.

Frequency distributions are often graphically represented by a histogram or frequency polygon. A histogram consists of a set of rectangles having bases on a horizontal axis (the *x* axis) with centres at the class marks and lengths equal to the class interval sizes and areas proportional to class frequencies. If the class intervals all have equal size, the heights of the rectangles are proportional to the class frequencies and it is then customary to take the heights numerically equal to the class frequencies. If class intervals do not have equal size, these heights must be adjusted. A frequency polygon is a line graph of class frequency plotted against class mark. It can be obtained by connecting midpoints of the tops of the rectangles in the histogram.

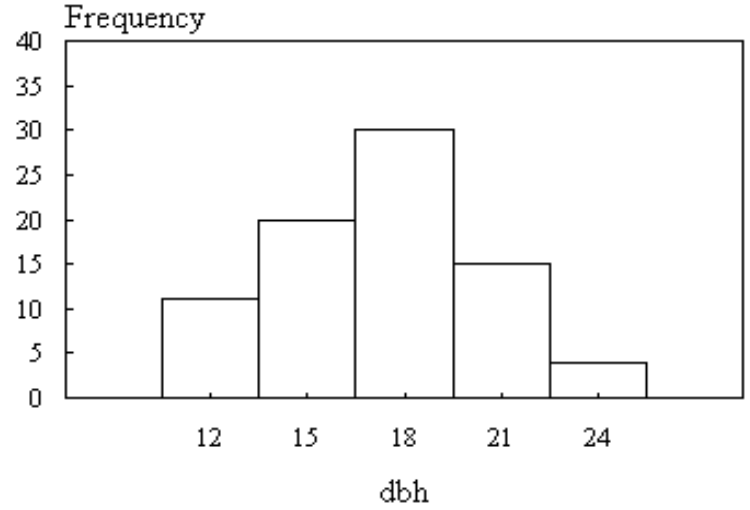


Figure 2.1. Histogram showing the frequency distribution of dbh

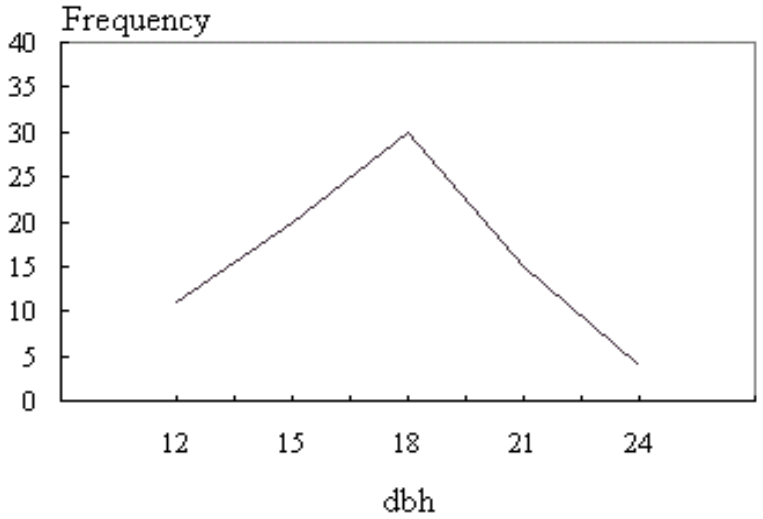


Figure 2.2. Frequency polygon showing the frequency distribution of dbh

2.3. Properties of frequency distribution

Having prepared a frequency distribution, a number of measures can be generated out of it, which leads to further condensation of the data. These are measures of location, dispersion, skewness and kurtosis.

2.3.1. Measures of location

A frequency distribution can be located by its average value which is typical or representative of the set of data. Since such typical values tend to lie centrally within a set of data arranged according to magnitude, averages are also called measures of central tendency. Several types of averages can be defined, the most common being the *arithmetic mean* or briefly the *mean*, *the median* and *the mode.* Each has advantages and disadvantages depending on the data and the intended purpose.

Arithmetic mean : The arithmetic mean or the mean of a set of N numbers $x_1, x_2, x_3, \ldots, x_N$ is denoted by $\bar{x}$ (read as 'x bar') and is defined as

$$\text{Mean} = \frac{x_1 + x_2 + x_3 + \ \ldots \ + x_N}{N} = \frac{\sum_{j=1}^{N} x_j}{N} = \frac{\sum x}{N} \tag{2.7}$$

The symbol $\sum_{j=1}^{N} x_j$ denote the sum of all the x_j's from $j = 1$ to $j = N$.

For example, the arithmetic mean of the numbers 8, 3, 5, 12, 10 is

$$\frac{8+3+5+12+10}{5} = \frac{38}{5} = 7.6$$

If the numbers $x_1, x_2, \ldots, x_K$ occur $f_1, f_2, \ldots, f_K$ times respectively (*i.e.*, occur with frequencies $f_1, f_2, \ldots, f_K$, the arithmetic mean is

$$\text{Mean} = \frac{f_1 x_1 + f_2 x_2 + \ldots + f_K x_K}{f_1 + f_2 + \ldots + f_K} = \frac{\sum_{j=1}^{K} f_j x_j}{\sum_{j=1}^{K} f_j} = \frac{\sum fx}{\sum f} \tag{2.8}$$

where $N = \sum f$ is the *total frequency*. *i.e.*, the total number of cases.

The computation of mean from grouped data of Table 2.1 is illustrated below.

Step 1. Find the midpoints of the classes. For this purpose add the lower and upper limits of the first class and divide by 2. For the subsequent classes go on adding the class interval.

Step 2. Multiply the midpoints of the classes by the corresponding frequencies, and add them up to get $\sum fx$.

The results in the above steps can be summarised as given in Table 2.2.

Table 2.2. Computation of mean from grouped data

Dbh class (cm)	Midpoint x	f	fx
11-13	12	11	132
14-16	15	20	300
17-19	18	30	540
20-22	21	15	315
23-25	24	4	96
Total		$\sum f = 80$	$\sum fx = 1383$

Step 3. Substitute the values in the formula

$$\text{Mean} = \frac{\sum fx}{\sum f}$$

$$= \frac{1383}{80} = 17.29 \text{ cm}$$

Median : The median of a set of numbers arranged in order of magnitude (*i.e.*, in an array) is the middle value or the arithmetic mean of the two middle values.

For example, the set of numbers 3, 4, 4, 5, 6, 8, 8, 8, 10 has median 6. The set of numbers 5, 5, 7, 9, 11, 12, 15, 18 has median $\frac{1}{2}(9+11) = 10$.

For grouped data the median, obtained by interpolation, is given by

$$\text{Median} = L_1 + \left(\frac{\left(\frac{N}{2} - \left(\sum f \right)_1 \right)}{f_m} \right) c \qquad (2.9)$$

where L_1 = lower class boundary of the median class (*i.e.*, the class containing the median)

N = number of items in the data (*i.e.*, total frequency)

$\left(\sum f\right)_1$ = sum of frequencies of all classes lower than the median class

f_m = frequency of median class

c = size of median class interval.

Geometrically, the median is the value of x (abscissa) corresponding to that vertical line which divides a histogram into two parts having equal areas.

The computation of median from grouped data of Table 2.1 is illustrated below.

Step 1. Find the midpoints of the classes. For this purpose add the lower and upper limits of the first class and divide by 2. For the subsequent classes go on adding the class interval.

Step 2. Write down the cumulative frequency and present the results as in Table 2.3.

Table 2.3. Computation of median from grouped data

Dbh class (cm)	Midpoint x	frequency f	Cumulative frequency
11-13	12	11	11
14-16	15	20	31
17-19	18	30	61
20-22	21	15	76
23-25	24	4	80
Total		$\sum f = 80$	

Step 3. Find the median class by locating the $(N / 2)$th item in the cumulative frequency column. In this example, $N / 2 = 40$. It falls in the class 17-19. Hence it is the median class.

Step 4. Use the formula (2.9) for calculating the median.

$$\text{Median} = 16.5 + \left(\frac{\left(\frac{80}{2} - 31\right)}{30}\right)3$$

$$= 17.4$$

Mode : The mode of a set of numbers is that value which occurs with the greatest frequency, *i.e.*, it is the most common value. The mode may not exist, and even if it does exist, it may not be unique.

The set of numbers 2, 2, 5, 7, 9, 9, 9, 10, 10, 11, 12, 18 has mode 9. The set 3, 5, 8, 10, 12, 15, 16 has no mode. The set 2, 3, 4, 4, 4, 5, 5, 7, 7, 7, 9 has two modes 4 and 7 and is called *bimodal.* A distribution having only one mode is called unimodal.

In the case of grouped data where a frequency curve have been constructed to fit the data, the mode will be the value (or values) of x corresponding to the maximum point (or points) on the curve.

From a frequency distribution or histogram, the mode can be obtained from the formula,

$$\text{Mode} = L_1 + \left(\frac{f_2}{f_1 + f_2}\right)c \qquad (2.10)$$

where L_1 = Lower class boundary of modal class (*i.e.*, the class containing the mode).
f_1 = Frequency of the class previous to the modal class.
f_2 = Frequency of the class just after the modal class.
c = Size of modal class interval.

The computation of mode from grouped data of Table 2.1. is illustrated below.

Step 1. Find out the modal class. The modal class is the class against the maximum frequency. In our example, the maximum frequency is 30 and hence the modal class is 17-19.

Step 2. Use the formula (2.10) for computing mode

$$\text{Mode} = 16.5 + \left(\frac{15}{15+20}\right)3$$

$$= 17.79$$

The general guidelines on the use of measures of location are that mean is mostly to be used in the case of symmetric distributions (explained in Section 2.3.3) as it is greatly affected by extreme values in the data, median has the distinct advantage of being computable even with open classes and mode is useful with multimodal distributions as it works out to be the most frequent observation in a data set.

2.3.2. Measures of dispersion

The degree to which numerical data tend to spread about an average value is called the variation or dispersion of the data. Various measures of dispersion or variation are available, like the range, mean deviation or semi-interquartile range but the most common is the standard deviation.

Standard deviation: The standard deviation of a set of N numbers $x_1, x_2, \ldots, x_N$ is defined by

$$\text{Standard deviation} = \sqrt{\frac{\sum_{j=1}^{N}\left(x_j - \bar{x}\right)^2}{N}} \qquad (2.11)$$

where $\bar{x}$ represents the arithmetic mean.

Thus standard deviation is the square root of the mean of the squares of the deviations of individual values from their mean or, as it is sometimes called, the *root mean square*

deviation. For computation of standard deviation, the following simpler form is used many times.

$$\text{Standard deviation} = \sqrt{\frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2} \qquad (2.12)$$

For example, the set of data given below represents diameters at breast-height of 10 randomly selected teak trees in a plot.

23.5, 11.3, 17.5, 16.7, 9.6, 10.6, 24.5, 21.0, 18.1, 20.7

Here $N = 10$, $\sum x^2 = 3266.5$ and $\sum x = 173.5$. Hence,

$$\text{Standard deviation} = \sqrt{\frac{3266.5}{10} - \left(\frac{173.5}{10}\right)^2} = 5.062$$

If $x_1, x_2, \ldots, x_K$ occur with frequencies $f_1, f_2, \ldots, f_K$ respectively, the standard deviation can be computed as

$$\text{Standard deviation} = \sqrt{\frac{\sum_{j=1}^{K} f_j\left(x_j - \bar{x}\right)^2}{N}} \qquad (2.13)$$

where $N = \sum_{j=1}^{K} f_j = \sum f$

Equation (2.13) can be written in the equivalent form which is useful in computations, as

$$\text{Standard deviation} = \sqrt{\frac{\sum fx^2}{N} - \left(\frac{\sum fx}{N}\right)^2} \qquad (2.14)$$

The *variance* of a set of data is defined as the square of the standard deviation. The ratio of standard deviation to mean expressed in percentage is called *coefficient of variation*.

For illustration, we can use the data given in Table 2.1.

Step 1. Find the midpoints of the classes. For this purpose, add the lower and upper limits of the first class and divide by 2. For the subsequent classes, go on adding the class interval.

Step 2. Multiply the midpoints of the classes by the corresponding frequencies, and add them up to get $\sum fx$.

Step 3. Multiply the square of the midpoints of the classes by the corresponding frequencies and add them up to get $\sum fx^2$.

The above results can be summarised as in Table 2.4.

Table 2.4. Computation of standard deviation from grouped data

Dbh class (cm)	Midpoint x	Frequency f	fx	fx^2
11-13	12	11	132	1584
14-16	15	20	300	4500
17-19	18	30	540	9720
20-22	21	15	315	6615
23-25	24	4	96	2304
Total		80	1383	24723

Step 4. Use the formula (2.14) for calculating the standard deviation and find out variance and coefficient of variation

$$\text{Standard deviation} = \sqrt{\frac{24723}{80} - \left(\frac{1383}{80}\right)^2} = 3.19$$

$$\text{Variance} = (\text{Standard deviation})^2 = (3.19)^2 = 10.18$$

$$\text{Coefficient of variation} = \frac{\text{Standard deviation}}{\text{Mean}}(100) = \frac{3.19}{17.29}(100) = 18.45$$

Both standard deviation and mean carry units of measurement where as coefficient of variation has no such units and hence is useful for comparing the extent of variation in characters which differ in their units of measurement. This is a useful property in comparison of variation in two sets of numbers which differ by their means. For instance, suppose that the variation in height of seedlings and that of older trees of a species are to be compared. Let the respective means and standard deviations be,

Mean height of seedlings = 50 cm, Standard deviation of height of seedlings = 10 cm.
Mean height of trees = 500 cm, Standard deviation of height of seedlings = 100 cm.

By the absolute value of the standard deviation, one may tend to judge that variation is more in the case of trees but the relative variation as indicated by the coefficient of variation (20 %) is the same in both the sets.

2.3.3. Measures of skewness

Skewness is the degree of asymmetry, or departure from symmetry, of a distribution. If the frequency curve (smoothed frequency polygon) of a distribution has a longer 'tail' to the right of the central maximum than to the left, the distribution is said to be *skewed*

to the right or to have *positive skewness*. If the reverse is true, it is said to be *skewed to the left* or to have *negative skewness*. An important measure of skewness expressed in dimensionless form is given by

$$\text{Moment coefficient of skewness} = \beta_1 = \frac{\mu_3^2}{\mu_2^3} \qquad (2.15)$$

where μ_2 and μ_3 are the second and third central moments defined using the formula,

$$\mu_r = \frac{\sum_{j=1}^{N}\left(x_j - \bar{x}\right)^r}{N} = \frac{\sum (x-\bar{x})^r}{N} \qquad (2.16)$$

For grouped data, the above moments are given by

$$\mu_r = \frac{\sum_{j=1}^{K} f_j\left(x_j - \bar{x}\right)^r}{N} = \frac{\sum f\,(x-\bar{x})^r}{N} \qquad (2.17)$$

For a symmetrical distribution, $\beta_1 = 0$. Skewness is positive or negative depending upon whether μ_3 is positive or negative.

The data given in Table 2.1 are used for illustrating the steps for computing the measure of skewness.

Step 1. Calculate the mean.

$$\text{Mean} = \frac{\sum fx}{\sum f} = 17.29$$

Step 2. Compute $f_j\,(x_j - \bar{x})^2$, $f_j\,(x_j - \bar{x})^3$ and their sum as summarised in Table 2.5.

Table 2.5. Steps for computing coefficient of skewness from grouped data

Dbh class (cm)	Midpoint x	f	$x_j - \bar{x}$	$f_j(x_j - \bar{x})^2$	$f_j(x_j - \bar{x})^3$	$f_j(x_j - \bar{x})^4$
11-13	12	11	-5.29	307.83	-1628.39	8614.21
14-16	15	20	-2.29	104.88	-240.18	550.01
17-19	18	30	0.71	15.12	10.74	7.62
20-22	21	15	3.71	206.46	765.97	2841.76
23-25	24	4	6.71	180.10	1208.45	8108.68
Total		80	3.55	814.39	116.58	20122.28

Step 3. Compute μ_2 and μ_3 using the formula (2.17).

$$\mu_2 = \frac{\sum f(x-\bar{x})^2}{N}$$

$$= \frac{814.39}{80}$$

$$= 10.18$$

$$\mu_3 = \frac{\sum f(x-\bar{x})^3}{N}$$

$$= \frac{116.58}{80}$$

$$= 1.46$$

Step 4. Compute the measure of skewness using the formula (2.15).

$$\text{Moment coefficient of skewness} = \beta_1 = \frac{(1.46)^2}{(10.18)^3}$$

$$= 0.002.$$

Since, $\beta_1 = .002$, the distribution is very slightly skewed or skewness is negligible. It is positively skewed since μ_3 is positive.

2.3.4. Kurtosis

Kurtosis is the degree of peakedness of a distribution, usually taken relative to a normal distribution. A distribution having a relatively high peak is called leptokurtic, while the curve which is flat-topped is called platykurtic. A bell shaped curve which is not very peaked or very flat-topped is called mesokurtic.

One measure of kurtosis, expressed in dimensionless form, is given by

$$\text{Moment coefficient of kurtosis} = \beta_2 = \frac{\mu_4}{\mu_2^2} \qquad (2.18)$$

where μ_4 and μ_2 can be obtained from the formula (2.16) for ungrouped data and by using the formula (2.17) for grouped data. The distribution is called normal if $\beta_2 = 3$. When β_2 is more than 3, the distribution is said to be leptokurtic. If β_2 is less than 3, the distribution is said to be platykurtic

For example, the data in Table 2.1 is utilised for computing the moment coefficient of kurtosis.

Step 1. Compute the mean .

$$\text{Mean} = \frac{\sum fx}{\sum f} = 17.29$$

Step 2. Compute $f_j (x_j - \bar{x})^2$, $f_j (x_j - \bar{x})^4$ and their sum as summarised in Table 2.5.

Step 3. Compute μ_2 and μ_4 using the formula (2.17).

$$\mu_2 = \frac{\sum f(x-\bar{x})^2}{N}$$

$$= \frac{814.39}{80}$$

$$= 10.18$$

$$\mu_4 = \frac{\sum f(x-\bar{x})^4}{N}$$

$$= \frac{20122.28}{80}$$

$$= 251.53$$

Step 4. Compute the measure of kurtosis using the formula (2.18).

$$\text{Moment coefficient of kurtosis} = \beta_2 = \frac{251.53}{(10.18)^2}$$

$$= 2.43.$$

The value of β_2 is 2.38 which is less than 3. Hence the distribution is platykurtic.

2.4. Discrete theoretical distributions

If a variable X can assume a discrete set of values $x_1, x_2, \ldots, x_K$ with respective probabilities $p_1, p_2, \ldots, p_K$ where $p_1 + p_2 + \ldots + p_K = 1$, we say that a *discrete probability distribution* for X has been defined. The function $p(x)$ which has the respective values $p_1, p_2, \ldots, p_K$ for $x = x_1, x_2, \ldots, x_K$, is called the *probability function* or *frequency function* of X. Because X can assume certain values with given probabilities, it is often called a *discrete random variable.*

For example, let a pair of fair dice be tossed and let X denote the sum of the points obtained. Then the probability distribution is given by the following table.

X	2	3	4	5	6	7	8	9	10	11	12
$p(x)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

The probability of getting sum 5 is $\frac{4}{36} = \frac{1}{4}$. Thus in 900 tosses of the dice, we would expect 100 tosses to give the sum 5.

Note that this is analogous to a relative frequency distribution with probabilities replacing relative frequencies. Thus we can think of probability distributions as theoretical or ideal limiting forms of relative frequency distributions when the number of observations is made very large. For this reason, we can think of probability distributions as being distributions for populations, whereas relative frequency distributions are distributions of samples drawn from this population.

When the values of x can be ordered as in the case where they are real numbers, we can define the cumulative distribution function,

$$F(x) = \sum_{z<x} p(z) \text{ for all } x \qquad (2.19)$$

$F(x)$ is the probability that X will take on some value less than or equal to x.

Two important discrete distributions which are encountered frequently in research investigations in forestry are mentioned here for purposes of future reference.

2.4.1. Binomial distribution

A binomial distribution arises from a set of n independent trials with outcome of a single trial being dichotomous such as 'success' or 'failure'. A binomial distribution applies if the probability of getting x successes out of n trials is given by the function,

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0,\ 1,\ 2,\ \ldots,\ n \qquad (2.20)$$

where n is a positive integer and $0<p<1$. The constants n and p are the parameters of the binomial distribution. As indicated, the value of x ranges from 0 to n.

For example, if a silviculturist is observing mortality of seedlings in plots in a plantation where 100 seedlings were planted in each plot and records live plants as 'successes' and dead plants as 'failures', then the variable 'number of live plants in a plot' may follow a binomial distribution.

Binomial distribution has mean np and a standard deviation $\sqrt{np(1-p)}$. The value of p is estimated from a sample by

$$\hat{p} = \frac{x}{n} \qquad (2.21)$$

where x is the number of successes in the sample and n is the total number of cases examined.

As an example, suppose that an entomologist picks up at random 5 plots each of size 10 m x 10 m from a plantation with seedlings planted at 2 m x 2 m espacement. Let the

observed number of plants affected by termites in the five plots containing 25 seedlings each be (4, 7, 7, 4, 3). The pooled estimate of *p* from the five plots would be,

$$\hat{p} = \frac{\sum x}{\sum n} = \frac{25}{125} = 0.2$$

Further, if he picks up a plot of same size at random from the plantation, the probability of that plot containing a specified number of plants infested with termites can be obtained by Equation (2.20) provided the infestation by termites follow binomial distribution. For instance, the probability of getting a plot uninfested by termites is

$$p(0) = \binom{25}{0} 0.2^0 (1 - 0.2)^{25}$$

$$= 0.0038$$

2.4.2. The Poisson distribution

A discrete random variable *X* is said to have a Poisson distribution if the probability of assuming specific value *x* is given by

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \qquad x = 0,\ 1,\ 2,\ \ldots \infty \tag{2.22}$$

where $\lambda > 0$. The variable *X* ranges from 0 to ∞.

In ecological studies, certain sparsely occurring organisms are found to be distributed randomly over space. In such instances, observations on number of organisms found in small sampling units are found to follow Poisson distribution. Poisson distribution has the single parameter λ which is the mean and also the variance of the distribution. Accordingly the standard deviation is $\sqrt{\lambda}$. From samples, the values of λ is estimated as

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.23}$$

where x_i's are the number of cases detected in a sampling unit and *n* is the number of sampling units observed.

For instance, a biologist observes the numbers of leech found in 100 samples taken from a fresh water lake. Let the total number of leeches caught be 80 so that the mean number per sample is calculated as,

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{80}{100} = 0.8$$

If the variable follows Poisson distribution, the probability of getting at least one leach in a fresh sample can be calculated as 1 - $p(0)$ which is,

$$1 - p(0) = 1 - \frac{(0.8)^0 e^{-0.8}}{0!}$$

$$= 0.5507$$

2.5. Continuous theoretical distributions

The idea of discrete distribution can be extended to the case where the variable X may assume continuous set of values. The relative frequency polygon of a sample becomes, in the theoretical or limiting case of a population, a continuous curve as shown in Figure 2.3, whose equation is $y = p(x)$.

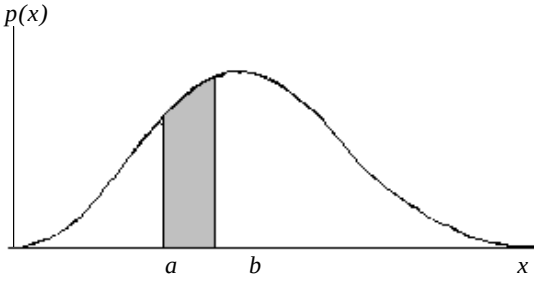


Figure 2.3. Graph of continuous distribution

The total area under this curve bounded by the X axis is equal to one, and the area under the curve between lines $X = a$ and $X = b$ (shaded in the figure) gives the probability that X lies between a and b, which can be denoted by $P(a<X<b)$. We call $p(x)$ a probability density function, or briefly a density function, and when such a function is given, we say that a continuous probability distribution for X has been defined. The variable X is then called a continuous random variable.

Cumulative distribution function for a continuous random variable is

$$F(x) = \int_{-\infty}^{x} f(t)dt \qquad (2.24)$$

The symbol $\int$ indicates integration which is in a way equivalent to summation in the discrete case. As in the discrete case, $F(x)$ gives the probability that the variable X will assume a value less than or equal to x. A useful property of the cumulative distribution function is that

$$P(a \le X \le b) = F(b) - F(a) \qquad (2.25)$$

Two cases of continuous theoretical distributions which frequently occur in forestry research are discussed here mainly for future references.

2.5.1. Normal distribution

Normal distribution is defined by the probability density function,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad -\infty < x, \mu < \infty \qquad 0 < \sigma \tag{2.26}$$

where μ is a location parameter and σ is a scale parameter. The range of the variable X is from $-\infty$ to $+\infty$. The μ parameter also varies from $-\infty$ to $+\infty$ but σ is always positive. The parameters μ and σ are not related. Equation (2.26) is a symmetrical function around μ as can be seen from Figure 2.4 which shows a normal curve for $\mu = 0$ and $\sigma = 1$. When $\mu = 0$ and $\sigma = 1$, the distribution is called a standard normal curve.

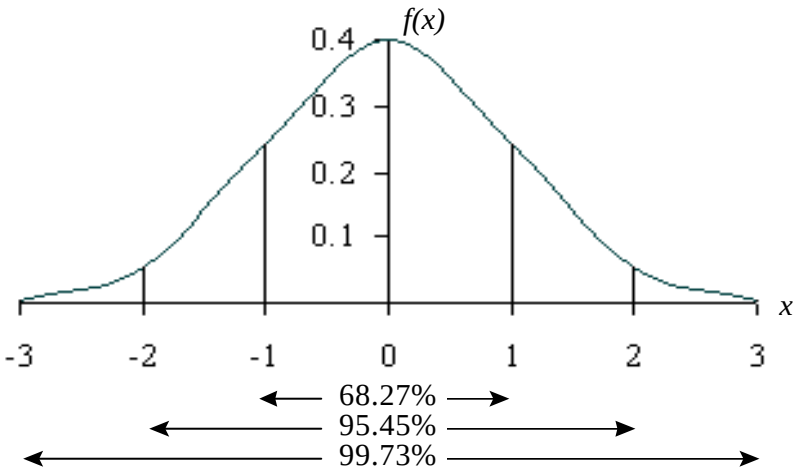


Figure 2.4. Graph of a normal distribution for $\mu = 0$ and $\sigma = 1$

If the total area bounded by the curve and the axis in Figure 2.4 is taken as unity, the area under the curve between two ordinates $X = a$ and $X = b$, where $a<b$, represents the probability that X lies between a and b, denoted by $P(a<X<b)$. Appendix 1 gives the areas under this curve which lies outside $+z$ and $-z$.

Normal distribution has mean μ and standard deviation σ. The distribution satisfies the following area properties. Taking the total area under the curve as unity, $\mu \pm \sigma$ covers 68.27% of the area, $\mu \pm 2\sigma$ covers 95.45% and $\mu \pm 3\sigma$ will cover 99.73 % of the total area. For instance, let the mean height of trees in a large plantation of a particular age be 10 m and the standard deviation be 1 m. Consider the deviation of height of individual trees from the population mean. If these deviations are normally distributed, we can expect about 68% of the trees to have their deviations from the mean within 1m; around 95% of the trees to have deviations lying with in 2 m and 99% of the trees showing deviations within 3 m.

Although normal distribution was originally proposed as a measurement error model, it was found to be basis of variation in a large number of biometrical characters. Normal distribution is supposed to arise from additive effects of a large number of independent causative random variables.

The estimates of μ and σ from sample observations are

$$\hat{\mu} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.27}$$

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} \tag{2.28}$$

where x_i, i = 1, …, n are n independent observations from the population.

2.5.2. Lognormal distribution

Let X be a random variable. Consider the transformation from X to Y by $Y = \ln X$. If the transformed variable Y is distributed according to a normal model, X is said to be a 'lognormal' random variable. The probability density function of lognormal distribution is given by

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2}, \qquad 0 < x, \sigma; \; -\infty < \mu < \infty \tag{2.29}$$

In this case, e^{μ} is a scale parameter and σ is a shape parameter. The shape of the log-normal distribution is highly flexible as can be seen from Figure 2.5 which plots equation (2.29) for different values of σ when $\mu = 0$.

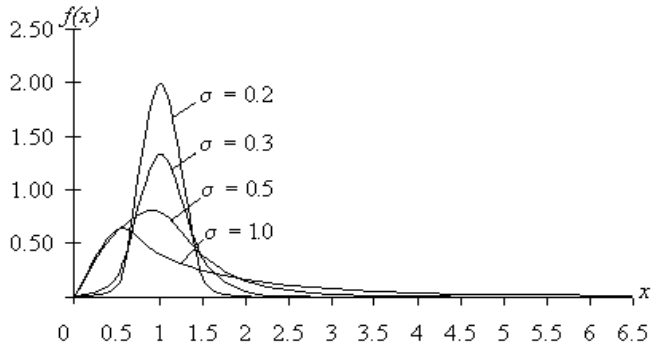


Figure 2.5. Graph of lognormal distribution for $\mu = 0$ and different values of σ.

The mean and standard deviation of log-normal distribution are complex functions of the parameters μ and σ. The mean and standard deviation are given by,

$$\text{Mean} = e^{\mu + \frac{\sigma^2}{2}} \tag{2.30}$$

$$\text{Standard deviation} = \sqrt{\left(e^{2\mu+\sigma^2}\right)\left(e^{\sigma^2}-1\right)} \tag{2.31}$$

Unlike the normal distribution, the mean and standard deviation of this distribution are not independent. This distribution also arises from compounding effects of a large number of independent effects with multiplicative effects rather than additive effects. For instance, if the data are obtained by pooling height of trees from plantations of differing age groups, it may show a log-normal distribution, the age having a compounding effect on variability among trees. Accordingly, trees of smaller age group may show low variation but trees of older age group are likely to exhibit large variation because of their interaction with the environment for a larger span of time.

For log-normal distribution, the estimates of the parameters of the μ and σ are obtained by

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \ln x_i \tag{2.32}$$

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\ln x_i - \hat{\mu}\right)^2} \tag{2.33}$$

where x_i, i = 1, …, n are n independent observations from the population.

More elaborate discussion including several solved problems and computational exercises on topics mentioned in this chapter can be found in Spiegel and Boxer (1972).

3 STATISTICAL INFERENCE

3.1. Tests of hypotheses

Any research investigation progresses through repeated formulation and testing of hypotheses regarding the phenomenon under consideration, in a cyclical manner. In order to reach an objective decision as to whether a particular hypothesis is confirmed by a set of data, we must have an objective procedure for either rejecting or accepting that hypothesis. Objectivity is emphasized because one of the requirements of the scientific method is that one should arrive at scientific conclusions by methods which are public and which may be repeated by other competent investigators. This objective procedure would be based on the information we obtain in our research and on the risk we are willing to take that our decision with respect to the hypothesis may be incorrect.

The general steps involved in testing hypotheses are the following. (i) Stating the null hypothesis (ii) Choosing a statistical test (with its associated statistical model) for testing the null hypothesis (iii) Specifying the significance level and a sample size (iv) Finding the sampling distribution of the test statistic under the null hypothesis (v) Defining the region of rejection (vi) Computing the value of test statistic using the data

obtained from the sample(s) and making a decision based on the value of the test statistic and the predefined region of rejection. An understanding of the rationale for each of these steps is essential to an understanding of the role of statistics in testing a research hypothesis which is discussed here with a real life example.

(i) Null hypothesis : The first step in the decision-making procedure is to state the null hypothesis usually denoted by H_0. The null hypothesis is a hypothesis of no difference. It is usually formulated for the express purpose of being rejected. If it is rejected, the alternative hypothesis H_1 may be accepted. The alternative hypothesis is the operational statement of the experimenter's research hypothesis. The research hypothesis is the prediction derived from the theory under test. When we want to make a decision about differences, we test H_0 against H_1. H_1 constitutes the assertion that is accepted if H_0 is rejected.

To present an example, suppose a forest manager suspects a decline in the productivity of forest plantations of a particular species in a management unit due to continued cropping with that species. This suspicion would form the research hypothesis. Confirmation of that guess would add support to the theory that continued plantation activity with the species in an area would lead to site deterioration. To test this research hypothesis, we state it in operational form as the alternative hypothesis, H_1. H_1 would be that the current productivity level for the species in the management unit (μ_1) is less than that of the past (μ_0). Symbolically, $\mu_1 < \mu_0$. The H_0 would be that $\mu_1 = \mu_0$. If the data permit us to reject H_0, then H_1 can be accepted, and this would support the research hypothesis and its underlying theory. The nature of the research hypothesis determines how H_1 should be stated. If the forest manager is not sure of the direction of change in the productivity level due to continued cropping, then H_1 is that $\mu_1 \neq \mu_0$.

(ii) The choice of the statistical test : The field of statistics has developed to the extent that we now have, for almost all research designs, alternative statistical tests which might be used in order to come to a decision about a hypothesis. The nature of the data collected largely determines the test criterion to be used. In the example considered here, let us assume that data on yield of timber on a unit area basis at a specified age can be obtained from a few recently felled plantations or parts of plantations of fairly similar size from the management unit. Based on the relevant statistical theory, a test statistic that can be chosen in this regard is,

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \qquad (3.1)$$

where $\bar{x}$ = Mean yield at a specified age from the recently felled plantations in the management unit.

σ = Standard deviation of the yield of the recently felled plantations in the management unit.

n = Number of recently felled plantations from which the data can be gathered.

μ_0 = Mean yield of plantations at the specified age in the management unit a few decades back based on a large number of past records.

The term 'statistic' refers to a value computed from the sample observations. The test statistic specified in Equation (3.1) is the deviation of the sample mean from the pre-

specified value , μ_0 , in relation to the variance of such deviations and the question is to what extent such deviations are permissible if the null hypothesis were to be true.

(iii) The level of significance and the sample size : When the null hypothesis and alternative hypothesis have been stated, and when the statistical test appropriate to the problem has been selected, the next step is to specify a level of significance (α) and to select a sample size (*n*). In brief, the decision making procedure is to reject H_0 in favour of H_1 if the statistical test yields a value whose associated probability of occurrence under H_0 is equal to or less than some small probability symbolized as α. That small probability is called the level of significance. Common values of α are 0.05 and 0.01. To repeat, if the probability associated with the occurrence under H_0, *i.e.*, when the null hypothesis is true, of the particular value yielded by a statistical test is equal to or less than α, we reject H_0 and accept H_1, the operational statement of the research hypothesis. It can be seen, then that α gives the probability of mistakenly or falsely rejecting H_0.

Since the value of α enters into the determination of whether H_0 is or is not rejected, the requirement of objectivity demands that α be set in advance of the collection of the data. The level at which the researcher chooses to set α should be determined by his estimate of the importance or possible practical significance of his findings. In the present example, the manager may well choose to set a rather stringent level of significance, if the dangers of rejecting the null hypothesis improperly (and therefore unjustifiably advocating or recommending a drastic change in management practices for the area) are great. In reporting his findings, the manager should indicate the actual probability level associated with his findings, so that the reader may use his own judgement in deciding whether or not the null hypothesis should be rejected.
There are two types of errors which may be made in arriving at a decision about H_0. The first, the *Type I error*, is to reject H_0 when in fact it is true. The second, the *Type II error*, is to accept H_0 when in fact it is false. The probability of committing a Type I error is given by α. The larger is α, the more likely it is that H_0 will be rejected falsely, *i.e.*, the more likely it is that Type I error will be committed. The Type II error is usually represented by β, *i.e.*, P(Type I error) = α, P(Type II error) = β. Ideally, the values of both α and β would be specified by the investigator before he began his investigations. These values would determine the size of the sample (*n*) he would have to draw for computing the statistical test he had chosen. Once α and *n* have been specified, β is determined. In as much as there is an inverse relation between the likelihood of making the two types of errors, a decrease in α will increase β for any given *n*. If we wish to reduce the possibility of both types of errors, we must increase *n*. The term 1 - β is called the power of a test which is the probability of rejecting H_0 when it is in fact false. For the present example, guided by certain theoretical reasons, let us fix the sample size as 30 plantations or parts of plantations of similar size drawn randomly from the possible set for gathering data on recently realized yield levels from the management unit.

(iv) The sampling distribution : When an investigator has chosen a certain statistical test to use with his data, he must next determine what is the sampling distribution of the

test statistic. It is that distribution we would get if we took all possible samples of the same size from the same population, drawing each randomly and workout a frequency distribution of the statistic computed from each sample. Another way of expressing this is to say that the sampling distribution is the distribution, under H_0, of all possible values that some statistic (say the sample mean) can take when that statistic is computed from randomly drawn samples of equal size. With reference to our example, if there were 100 plantations of some particular age available for felling, 30 plantations can be drawn randomly in $\binom{100}{30} = 2.937 \text{ x } 10^{25}$ ways. From each sample of 30 plantation units, we can compute a *z* statistic as given in Equation (3.1). A relative frequency distribution prepared using specified class intervals for the *z* values would constitute the sampling distribution of our test statistic in this case. Thus the sampling distribution of a statistic shows the probability under H_0 associated with various possible numerical values of the statistic. The probability associated with the occurrence of a particular value of the statistic under H_0 is not the probability of just that value rather, the probability associated with the occurrence under H_0 of a particular value plus the probabilities of all more extreme possible values. That is, the probability associated with the occurrence under H_0 of a value as extreme as or more extreme than the particular value of the test statistic.

It is obvious that it would be essentially impossible for us to generate the actual sampling distribution in the case of our example and ascertain the probability of obtaining specified values from such a distribution. This being the case, we rely on the authority of statements of proved mathematical theorems. These theorems invariably involve assumptions and in applying the theorems we must keep the assumptions in mind. In the present case it can be shown that the sampling distribution of *z* is a normal distribution with mean zero and standard deviation unity for large sample size (*n*). When a variable is normally distributed, its distribution is completely characterised by the mean and the standard deviation. This being the case, the probability that an observed value of such a variable will exceed any specified value can be determined. It should be clear from this discussion and this example that by knowing the sampling distribution of some statistic we are able to make probability statements about the occurrence of certain numerical values of that statistic. The following sections will show how we use such a probability statement in making a decision about H_0.

(v) The region of rejection : The sampling distribution includes all possible values a test statistic can take under H_0. The region of rejection consists of a subset of these possible values, and is defined so that the probability under H_0 of the occurrence of a test statistic having a value which is in that subset is α. In other words, the region of rejection consists of a set of possible values which are so extreme that when H_0 is true, the probability is very small (*i.e.*, the probability is α) that the sample we actually observe will yield a value which is among them. The probability associated with any value in the region of rejection is equal to or less than α.

The location of the region of rejection is affected by the nature of H_1. If H_1 indicates the predicted direction of the difference, then a one-tailed test is called for. If H_1 does not

indicate the direction of the predicted difference, then a two-tailed test is called for. One-tailed and two-tailed tests differ in the location (but not in the size) of the region of rejection. That is, in one-tailed test, the region of rejection is entirely at one end (one tail) of the sampling distribution. In a two-tailed test, the region of rejection is located at both ends of the sampling distribution. In our example, if the manager feels that the productivity of the plantations will either be stable or only decline over time, then the test he would carry out will be one-tailed. If the manager is uncertain about the direction of change, it will be the case for a two-tailed test.

The size of the region is expressed by α, the level of significance. If $\alpha = 0.05$, then the size of the region of rejection is 5 per cent of the entire space included under the curve in the sampling distribution. One-tailed and two-tailed regions of rejection for $\alpha = 0.05$ are illustrated in Figure 3.1. The regions differ in location but not in total size.

(vi) The decision : If the statistical test yields a value which is in the region of rejection, we reject H_0. The reasoning behind this decision process is very simple. If the probability associated with the occurrence under the null hypothesis of a particular value in the sampling distribution is very small, we may explain the actual occurrence of that value in two ways: first, we may explain it by deciding that the null hypothesis is false, or second, we may explain it by deciding that a rare and unlikely event has occurred. In the decision process, we choose the first of these explanations. Occasionally, of course, the second may be the correct one. In fact, the probability that the second explanation is the correct one is given by α, for rejecting H_0 when in fact it is true is the Type I error.

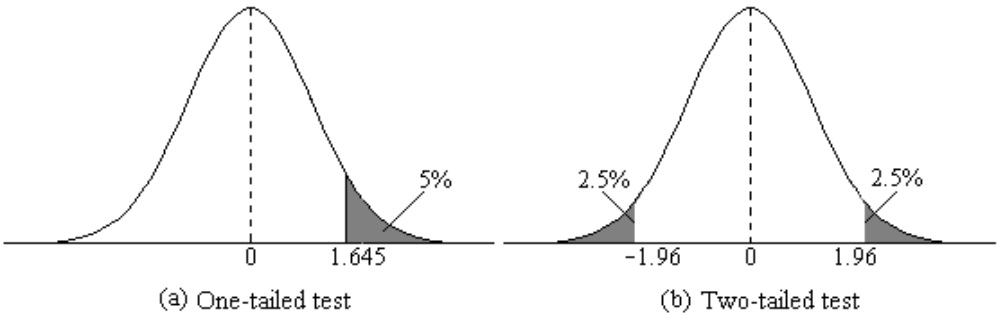


Figure 3.1. Sampling distribution of *z* under H_0 and regions of rejection for one-tailed and two-tailed tests.

When the probability associated with an observed value of a statistical test is equal to or less than the previously determined value of α, we conclude that H_0 is false. Such an observed value is called *significant*. H_0 , the hypothesis under test, is rejected whenever a significant result occurs. A significant value is one whose associated probability of occurrence under H_0 is equal to or less than α.

Coming back to our example, suppose that the mean timber yield obtained from 30 recently felled plantations at the age of 50 years in a particular management unit is 93 m^3/ha with a standard deviation of 10 m^3/ha. If the past records had revealed that the

mean yield realized from the same management unit a few decades back was 100 m^3/ha at comparable age, the value of the test statistic in our case would be

$$z = \frac{\bar{x} - m_0}{s\ /\sqrt{n}} = \frac{93 - 100}{10 / \sqrt{30}} = -3.834$$

Reference to Appendix 1 would show that the probability of getting such a value if the H_0 were to be true is much less than 0.05 taken as the prefixed level of significance. Hence the decision would be to accept the alternative hypothesis that there has been significant decline in the productivity of the management unit with respect to the plantations of the species considered.

The reader who wishes to gain a more comprehensive understanding of the topics explained in this section may refer Dixon and Massey (1951) for an unusually clear introductory discussion of the two types of errors, and to Anderson and Bancroft (1952) or Mood (1950) for advanced discussions of the theory of testing hypotheses. In the following sections, procedures for testing certain specific types of hypotheses are described.

3.2 Test of difference between means

It is often desired to compare means of two groups of observations representing different populations to find out whether the populations differ with respect to their locations. The null hypothesis in such cases will be 'there is no difference between the means of the two populations'. Symbolically, $H_0: \mu_1 = \mu_2$. The alternative hypothesis is $H_1: \mu_1 \neq \mu_2$ *i.e.*, $\mu_1 < \mu_2$ or $\mu_1 > \mu_2$.

3.2.1. Independent samples

For testing the above hypothesis, we make random samples from each population. The mean and standard deviation for each sample are then computed. Let us denote the mean as $\bar{x}_1$ and standard deviation as s_1 for the sample of size n_1 from the first population and the mean as $\bar{x}_2$ and standard deviation as s_2 for the sample of size n_2 from the second population. A test statistic that can be used in this context is,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad (3.2)$$

where $\bar{x}_1 = \frac{\sum x_{1i}}{n_1}$, $\bar{x}_2 = \frac{\sum x_{2i}}{n_2}$

s^2 is the pooled variance given by

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$s_1^2 = \frac{\sum x_{1i}^2 - \frac{\left(\sum x_{1i}\right)^2}{n_1}}{n_1 - 1} \quad \text{and} \quad s_2^2 = \frac{\sum x_{2i}^2 - \frac{\left(\sum x_{2i}\right)^2}{n_2}}{n_2 - 1}$$

The test statistic t follows Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom. The degree of freedom in this particular case is a parameter associated with the t distribution which governs the shape of the distribution. Although the concept of degrees of freedom is quite abstruse mathematically, generally it can be taken as the number of independent observations in a data set or the number of independent contrasts (comparisons) one can make on a set of parameters.

This test statistic is used under certain assumptions *viz*., (i) The variables involved are continuous (ii) The population from which the samples are drawn follow normal distribution (iii) The samples are drawn independently (iv) The variances of the two populations from which the samples are drawn are homogeneous (equal). The homogeneity of two variances can be tested by using *F*-test described in Section 3.3.

As an illustration, consider an experiment set up to evaluate the effect of inoculation with mycorrhiza on the height growth of seedlings of *Pinus kesiya*. In the experiment, 10 seedlings designated as Group I were inoculated with mycorrhiza while another 10 seedlings (designated as Group II) were left without inoculation with the microorganism. Table 3.1 gives the height of seedlings obtained under the two groups of seedlings.

Table 3.1. Height of seedlings of *Pinus kesiya* belonging to the two groups

Plot	Group I	Group II
1	23.0	8.5
2	17.4	9.6
3	17.0	7.7
4	20.5	10.1
5	22.7	9.7
6	24.0	13.2
7	22.5	10.3
8	22.7	9.1
9	19.4	10.5
10	18.8	7.4

Under the assumption of equality of variance of seedling height in the two groups, the analysis can proceed as follows.

Step1. Compute the means and pooled variance of the two groups of height measurements using the corresponding formulae as shown in Equation (3.2).

$\bar{x}_1 = 20.8$, $\bar{x}_2 = 9.61$

$$s_1^2 = \frac{(23.0)^2 + (17.4)^2 + \ldots + (18.8)^2 - \frac{(208)^2}{10}}{10-1}$$

$$= \frac{57.24}{9} = 6.36$$

$$s_2^2 = \frac{(8.5)^2 + (9.6)^2 + \ldots + (7.4)^2 - \frac{(96.1)^2}{10}}{10-1}$$

$$= \frac{24.3}{9} = 2.7$$

$$s^2 = \frac{(10-1)(6.36) + (10-1)(2.7)}{10+10-2}$$

$$= \frac{57.24 + 24.43}{18}$$

$$= 4.5372$$

Step 2. Compute the value of *t* using Equation (3.2)

$$t = \frac{20.8 - 9.61}{\sqrt{4.5372\left(\frac{1}{10} + \frac{1}{10}\right)}}$$

$$= 11.75$$

Step 3. Compare the computed value of t with the tabular value of t at the desired level of probability for $n_1 + n_2 - 2 = 18$ degrees of freedom.

Since we are not sure of the direction of the effect of mycorrhiza on the growth of seedlings, we may use a two-tailed test in this case. Referring Appendix 2, the critical values are -2.10 and +2.10 on either side of the distribution. For our example the computed value of t (11.75) is greater than 2.10 and so we may conclude that the populations of inoculated and uninoculated seedlings represented by our samples are significantly different with respect to their mean height.

The above procedure is not applicable if the variances of the two populations are not equal. In such cases, a slightly different procedure is followed and is given below:

Step 1. Compute the value of test statistic *t* using the following formula,

$$t = \frac{\left(\bar{x}_1 - \bar{x}_2\right)}{\sqrt{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]}} \tag{3.3}$$

Step 2. Compare the computed *t* value with a weighted tabular *t* value (*t'*) at the desired level of probability. The weighted tabular *t* value is computed as shown below.

$$t' = \frac{w_1 t_1 + w_2 t_2}{w_1 + w_2} \tag{3.4}$$

where $w_1 = \frac{s_1^2}{n_1}$, $w_2 = \frac{s_2^2}{n_2}$,

t_1 and t_2 are the tabular values of Student's *t* at $\left(n_1 - 1\right)$ and $\left(n_2 - 1\right)$ degrees of freedom respectively, at the chosen level of probability.

For example, consider the data given in Table 3.1. The homogeneity of variances of the two groups can be tested by using *F*-test given in Section 3.3. In case the two variances are not equal, the test statistic *t* will be computed as,

$$t = \frac{(20.8 - 9.61)}{\sqrt{\left[\frac{6.36}{10} + \frac{2.7}{10}\right]}} = 11.76$$

$$t' = \frac{(0.636)(2.26) + (0.270)(2.26)}{0.636 + 0.270} = 2.26$$

Since the computed *t* value (11.76) is greater than the tabular value (2.26), we may conclude that the two means are significantly different. Here, the value of *t'* remained the same as t_1 and t_2 because n_1 and n_2 are the same. This need not be the case always.

3.2.2. Paired samples

While comparing the means of two groups of observations, there could be instances where the groups are not independent but paired such as in the comparison of the status of a set of individuals before and after the execution of a treatment, in the comparison of say, the properties of bottom and top portion of a set of cane stems etc. In such situations, two sets of observations come from a single set of experimental units. Pairing of observations could occur on other grounds as well such as the case of pairs

of stem cuttings obtained from different mother plants and the individuals of a pair subjected to two different treatments with the objective of comparing the effect of the two treatments on the cuttings. The point to be noted is that the observations obtained from such pairs could be correlated. The statistical test used for comparing means of paired samples is generally called paired *t*- test.

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, be the n paired observations. Let the observations on x variable arise from a population with mean μ_1 and the observations on y variable arise from a population with mean μ_2. The hypothesis to be tested is $H_0: \mu_1 = \mu_2$. If we form $d_i = x_i - y_i$ for $i = 1, 2, \ldots, n$, which can be considered from a normal population with mean zero and known variance, the test statistic could be,

$$t = \frac{\bar{d}}{\sqrt{\frac{s_d^2}{n}}} \qquad (3.5)$$

where $s_d^2 = \frac{1}{n-1}\left(\sum d_i^2 - \frac{\left(\sum d_i\right)^2}{n}\right)$

The test statistic t in Equation (3.5) follows a Student's t distribution with $n-1$ degrees of freedom. The computed value of t is then comparable with the tabular value of t for $n-1$ degrees of freedom, at the desired level of probability.

For example, consider the data given in Table 3.2, obtained from soil core samples drawn from two different depth levels in a natural forest. The data pertain to organic carbon content measured at two different layers of a number of soil pits and so the observations are paired by soil pits. The paired *t*-test can be used in this case to compare the organic carbon status of soil at the two depth levels. The statistical comparison would proceed as follows.

Step 1. Get the difference between each pair of observations as shown in Table 3.2

Table 3.2. Organic carbon content measured from two layers of a set of soil pits from natural forest.

Soil pit	Organic carbon (%)		
	Layer 1 (x)	Layer 2 (y)	Difference (d)
1	1.59	1.21	0.38
2	1.39	0.92	0.47
3	1.64	1.31	0.33
4	1.17	1.52	-0.35
5	1.27	1.62	-0.35
6	1.58	0.91	0.67
7	1.64	1.23	0.41
8	1.53	1.21	0.32

9	1.21	1.58	-0.37
10	1.48	1.18	0.30

Step 2. Calculate mean difference and variance of the differences as shown in Equation (3.5).

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{1.81}{10} = 0.181$$

$$s_d^2 = \frac{1}{10-1}\left(\left((0.38)^2 + (0.47)^2 + \ . \ . \ .+(0.30)^2\right) - \frac{(1.81)^2}{10} \right)$$

$$= \frac{1.33789}{9} = 0.1486$$

Step 3. Calculate the value for t by substituting the values of $\bar{d}$ and s_d^2 in Equation (3.5).

$$t = \frac{0.181}{\sqrt{\frac{0.1486}{10}}} = 1.485$$

The value we have calculated for t (1.485) is less than the tabular value, 2.262, for 9 degrees of freedom at the 5% level of significance. It may therefore be concluded that there is no significant difference between the mean organic carbon content of the two layers of soil.

3.3. Test of difference between variances

We often need to test whether two independent random samples come from populations with same variance. Suppose that first sample of n_1 observations has a sample variance s_1^2 and that second sample of n_2 observations has a sample variance s_2^2 and that both samples come from normal distributions. The null hypothesis to be tested is that the two samples are independent random samples from normal populations with the same variance. Symbolically,

$$H_0: \sigma_1^2 = \sigma_2^2$$

where σ_1^2, σ_2^2 are populations variances of two populations from which the two samples are taken. The alternative hypothesis is

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

The test statistic used to test the above null hypothesis is

$$F = \frac{s_1^2}{s_2^2} \tag{3.6}$$

where s_1^2 is the larger mean square.

Under the null hypothesis, the test statistic may be shown to follow an F distribution with $(n_1 - 1, n_2 - 1)$ degrees of freedom. The decision rule is that if the calculated value of the test statistic is smaller than the critical value of the *F*-distribution at the desired probability level, we accept the null hypothesis that two samples are taken from populations having same variance, otherwise we reject the null hypothesis.

For example, let the variance estimates from two populations be $s_1^2 = 21.87$ and $s_2^2 = 15.36$ based on n_1=11and n_2= 8, observations from the two populations respectively. For testing the equality of population variances, we compute,

$$F = \frac{s_1^2}{s_2^2} = \frac{21.87}{15.36} = 1.424$$

and compare with the critical value of *F* distribution for 10 and 7 degrees of freedom. Referring Appendix 3, the critical value of *F* is 3.14 at the probability level of .05. Here the calculated value is less than the critical value and hence we conclude that the variances are equal.

3.4 Test of proportions

When the observations form counts belonging to particular categories such as 'diseased' or 'healthy', 'dead' or 'alive' etc. the data are usually summarized in terms of proportions. We may then be interested in comparing the proportions of incidence of an attribute in two populations. The null hypothesis set up in such cases is $H_0: P_1 = P_2$ and the alternative hypothesis is $H_1: P_1 \neq P_2$ (or $P_1 > P_2$ or $P_1 < P_2$) where P_1 and P_2 are proportions representing the two populations. In order to test our hypothesis, we take two independent samples of large size, say n_1 and n_2 from the two populations and obtain two sample proportions p_1 and p_2, respectively. The test statistic used is,

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \tag{3.7}$$

where q_1 = 1 - p_1, q_2 = 1 - p_2, This statistic follows a standard normal distribution.

As an example, consider an experiment on rooting of stem cuttings of *Casuarina equisetifolia* wherein the effect of dipping the cuttings in solutions of IBA at two different concentrations was observed. Two batches of 30 cuttings each, were subjected dipping treatment at concentrations of 50 and 100 ppm of IBA solutions respectively. Based on the observations on number of cuttings rooted in each batch of 30 cuttings, the following proportions of rooted cuttings under each concentration were obtained. At 50 ppm, the proportion of rooted cuttings was 0.5 and at 100 ppm, the proportion was 0.37. The question of interest is whether the observed proportions are indicative of significant differences in the effect of IBA at the two concentrations.

In accordance with our notation, here, $p_1 = 0.5$ and $p_2 = 0.37$. Then $q_1 = 0.5$, $q_2 = 0.63$. The value of $n_1 = n_2 = 30$. The value of the test statistic is,

$$z = \frac{0.5 - 0.37}{\sqrt{\frac{(0.5)(0.5)}{30} + \frac{(0.37)(0.63)}{30}}} = 1.024$$

Since the calculated value of z (1.024) is less than the table value (1.96) at 5% level of significance, we can conclude that there is no significant difference between proportion rooted cuttings under the two concentration levels.

3.5 Test of goodness of fit

In testing of hypothesis, sometimes our objective may be to test whether a sample has come from a population with a specified probability distribution. The expected distribution may one based on theoretical distributions like the normal, binomial or Poisson or a pattern expected under technical grounds. For instance, one may be interested in testing whether a variable like the height of trees follows normal distribution. A tree breeder may be interested to know whether the observed segregation ratios for a character deviate significantly from the Mendelian ratios. In such situations, we want to test the agreement between the observed and theoretical frequencies. Such a test is called a test of goodness of fit.

For applying the goodness of fit test, we use only the actual observed frequencies and not the percentages or ratios. Further, the observations within the sample should be non-overlapping and thereby independent. The expected frequency in each category should preferably be more than 5. The total number of observations should be large, say, more than 50.

The null hypothesis in goodness of fit tests is that there is no disagreement between the observed and theoretical distributions, or the observed distribution fits well with the theoretical distribution. The test statistic used is,

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \qquad (3.8)$$

where O_i = Observed frequency in the ith class,
E_i = Expected frequency in the ith class.
k = Number of categories or classes.

The χ^2 statistic of Equation (3.8) follows a χ^2-distribution with k-1 degrees of freedom. In case the expected frequencies are derived from parameters estimated from the sample, the degrees of freedom is (k-p-1) (where p is the number of parameters estimated). For example, in testing the normality of a distribution μ and σ^2 would be estimated from the sample by $\bar{x}$ and $s^2$and the degrees of freedom would therefore reduce to (k-2-1).

The expected frequencies may be computed based on the probability function of the appropriate theoretical distribution as relevant to the situation or it may be derived based on the scientific theory being tested like Mendel's law of inheritance. In the absence of a well defined theory, we may assume that all the classes are equally frequent in the population. For example, the number of insects caught in a trap in different times of a day, frequency of sighting an animal in different habitats etc. may be expected to be equal initially and subjected to the statistical testing. In such cases, the expected frequency is computed as

$$E = \frac{\text{Total of the observed frequencies}}{\text{Number of groups}} = \frac{n}{k} \qquad (3.9)$$

For example, consider the data given in Table 3.3. which represents the number of insect species collected from an undisturbed area at Parambikkulam Wildlife Sanctuary in different months. To test whether there are any significant differences between the number of insect species found in different months, we may state the null hypothesis as the diversity in terms of number of insect species is the same in all months in the sanctuary and derive the expected frequencies in different months accordingly.

Table 3.3. Computation of χ^2 using the data on number of species of insects collected from Prambikkulam in different months.

Month	O	E	$(O-E)^2/E$
Jan.	67	67	0.00
Feb.	115	67	34.39
Mar.	118	67	38.82
Apr.	72	67	0.37
May	67	67	0.00
Jun.	77	67	1.49
Jul.	75	67	0.96
Aug.	63	67	0.24
Sep.	42	67	9.33
Oct.	24	67	27.60

Nov.	32	67	18.28
Dec.	52	67	3.36
Total	804	804	134.84

The calculated χ^2 value is 134.84. Entering the χ^2 table (Appendix 4) for (12-1) = 11 degrees of freedom and $\alpha = 0.05$, we get the critical value of χ^2 as 19.7. Therefore, we accept the null hypothesis and conclude that the occurrence of the number of insect species in different months is the same.

3.6. Analysis of variance

Analysis of variance (ANOVA) is basically a technique of partitioning the overall variation in the responses observed in an investigation into different assignable sources of variation, some of which are specifiable and others unknown. Further, it helps in testing whether the variation due to any particular component is significant as compared to residual variation that can occur among the observational units.

Analysis of variance proceeds with an underlying model which expresses the response as a sum of different effects. As an example, consider Equation (3.10).

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1, 2, \ldots, t; \quad j = 1, 2, \ldots, n_i \tag{3.10}$$

where y_{ij} is the response of the jth individual unit belonging to the ith category or group, μ is overall population mean, α_i is the effect of being in the ith group and e_{ij} is a random error attached to the (ij)th observation. This constitutes a one-way analysis of variance model which can be expanded further by adding more and more effects as applicable to a particular situation. When more than one known source of variation is involved, the model is referred as multi-way analysis of variance model.

In order to perform the analysis, certain basic assumptions are made about the observations and effects. These are (i) The different component effects are additive (ii) The errors e_{ij} are independently and identically distributed with mean zero and constant variance.

Model (3.10) can also be written as

$$y_{ij} = \mu_i + e_{ij} \tag{3.11}$$

where $\mu_i = \mu + \alpha_i$

Under certain additional assumptions, analysis of variance also leads testing the following hypotheses that

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_t$$
$$H_1: \mu_i \neq \mu_j \quad \text{for at least one } i \text{ and } j \tag{3.12}$$

The additional assumption required is that the errors are distributed normally. The interpretation of the analysis of variance is valid only when such assumptions are met although slight deviations from these assumptions do not cause much harm.

Certain additional points to be noted are that the effects included in the model can be either fixed or random. For example, the effects of two well defined levels of irrigation are fixed as each irrigation level can be reasonably taken to have a fixed effect. On the other hand, if a set of provenances are randomly chosen from a wider set possible, the effects due to provenances is taken as random. The random effects can belong to a finite or an infinite population. The error effects are always random and may belong either to a finite or infinite population. A model in which all the effects are fixed except the error effect which is always taken as random is called a fixed effects model. Models in which both fixed and random effects occur are called mixed models. Models wherein all the effects are random are called random effects models. In fixed effects models, the main objectives will be to estimate the fixed effects, to quantify the variation due to them in the response and finally find the variation among the error effects. In random effects models, the emphasis will be on estimating the variation in each category of random effects. The methodology for obtaining expressions of variability is mostly the same in the different models, though the methods for testing are different.

The technique of analysis of variance is illustrated in the following with a one-way model involving only fixed effects. More complicated cases are dealt with in the Chapters 4 and 6 while illustrating the analyses related to different experimental designs.

3.6.1. Analysis of one-way classified data.

Consider a set of observations on wood density obtained on a randomly collected set of stems belonging to a set of cane species. Let there be t species with r observations coming from each species. The results may be tabulated as shown in the following table.

	Species					
	1	2	. .	i	. .	t	
	y_{11}	y_{21}		y_{i1}		y_{t1}	
	y_{12}	y_{22}		y_{i2}		y_{t2}	
	. .						
	y_{1j}	y_{2j}		y_{ij}		y_{tj}	
	. .						
	y_{1r}	y_{2r}		y_{ir}		y_{tr}	
Total	$y_{1.}$	$y_{2.}$		$y_{i.}$		$y_{t.}$	$y_{..}$= Grand total
Mean	$\bar{y}_1$	$\bar{y}_2$		$\bar{y}_i$		$\bar{y}_t$	$\bar{y}$ = Grand mean

Note: In the above table, a period (.) in the subscript denotes sum over that subscript.

The theory behind the analysis of variance would look complex for a nonmathematical reader. Hence a heuristic derivation of the formulae is given here. Consider the *r* observations which belong to any particular species, say *i*th species. Their values may not be the same. This demonstrates the influences of many extraneous factors operating on the observations on stems of that species. This influence may be measured in terms of the deviations of the individual observations from their mean value. Squared deviations would be better as raw deviations are likely to cancel out while summing up. Thus the extent of random variation incident on the observations on the *i*th species is given by

$$\left(y_{i1}-\bar{y}_{i.}\right)^2+\left(y_{i2}-\bar{y}_{i.}\right)^2+\ \ldots\ +\left(y_{ir}-\bar{y}_{i.}\right)^2\ =\sum_{j=1}^{r}\left(y_{ij}-\bar{y}_{i.}\right)^2 \qquad (3.13)$$

The variation produced by the external sources in the case of each species is a reflection of the influence of the uncontrolled factors and we can obtain a pooled estimate of their influence by their sum. Thus the total variability observed due to extraneous factors also generally known as sum of squares due to errors (*SSE*) is given by

$$SSE\ =\sum_{i=1}^{t}\sum_{j=1}^{r}\left(y_{ij}-\bar{y}_{i.}\right)^2 \qquad (3.14)$$

Besides random fluctuations, different species may carry different effects on the mean response. Thus variability due to *i*th species reflected in the *r* observations is

$$r\left(\bar{y}_{i.}-\bar{y}\right)^2 \qquad (3.15)$$

Thus the variability due to differences between the species is given by

$$SS \text{ due to species} = SSS = r\sum_{i=1}^{t}\left(y_{i.}-\bar{y}\right)^2 \qquad (3.16)$$

which can be shown algebraically equivalent to

$$SSS=\frac{\sum_{i=1}^{t}y_{i.}^2}{r}-\frac{\left(\sum_{i=1}^{t}\sum_{j=1}^{r}y_{ij}\right)^2}{tr} \qquad (3.17)$$

The second term of Equation (3.17) is called the correction factor (*C.F.*).

$$C.F.=\frac{\left(\sum_{i=1}^{t}\sum_{j=1}^{r}y_{ij}\right)^2}{tr} \qquad (3.18)$$

Finally, we have to find out the total variability present in all observations. This is given by the sum of the squares of deviations of all the responses from their general mean. It is given by

$$SSTO\ =\sum_{i=1}^{t}\sum_{j=1}^{r}\left(y_{ij}-\bar{y}\right)^2 \qquad (3.19)$$

$$= \sum_{i=1}^{t}\sum_{j=1}^{r}\left(y_{ij} - \bar{y}_{i.} + \bar{y}_{i.} - \bar{y}\right)^2$$

$$= \sum_{i=1}^{t}\sum_{j=1}^{r}\left(\left(y_{ij} - \bar{y}_{i.}\right)^2 + \left(\bar{y}_{i.} - \bar{y}\right)^2 + 2\left(y_{ij} - \bar{y}_{i.}\right)\left(\bar{y}_{i.} - \bar{y}\right)\right)$$

$$= \sum_{i=1}^{t}\sum_{j=1}^{r}\left(y_{ij} - \bar{y}_{i.}\right)^2 + r\sum_{i=1}^{t}\left(\bar{y}_{i.} - \bar{y}\right)^2 \qquad (3.20)$$

where $\sum_{i=1}^{t}\sum_{j=1}^{r} 2\left(y_{ij} - \bar{y}_{i.}\right)\left(\bar{y}_{i.} - \bar{y}\right) = 2\sum_{i=1}^{t}\left(\bar{y}_{i.} - \bar{y}\right)\sum_{j=1}^{r}\left(y_{ij} - \bar{y}_{i.}\right) = 0$

Thus the total variability in the responses could be expressed as a sum of variation between species and variation within species, which is the essence of analysis of variance.

For computational purposes, the *SSTO* can also be obtained as

$$SSTO = \sum_{i=1}^{t}\sum_{j=1}^{r}\left(y_{ij} - \bar{y}_{i.}\right)^2 + r\sum_{i=1}^{t}\left(\bar{y}_{i.} - \bar{y}\right)^2 = \sum_{i=1}^{t}\sum_{j=1}^{r} y_{ij}^2 - \frac{\left(\sum_{i=1}^{t}\sum_{j=1}^{r} y_{ij}\right)^2}{tr} \qquad (3.21)$$

The partitioning of total variability, as due to species differences and as due to extraneous factors though informative, are by themselves not very useful for further interpretation. This is because, their values depend on the number of species and the number of observations made on each species. In order to eliminate this effect of number of observations, the observed variability measures are reduced to variability per observation, *i.e.*, mean sum of squares. Since there are *rt* observations in all, yielding the total sum of squares, the mean sum of squares can of course be calculated by dividing total sum of squares by *rt*. Instead, this is divided by (*rt*-1), which is the total number of observations less one. This divisor is called the *degrees of freedom* which indicate the number of independent deviations from the mean contributing to the computation of total variation. Hence,

Mean sum of squares due to species = $MSS = \dfrac{SSS}{t-1}$ (3.22)

Mean sum of squares due to error = $MSE = \dfrac{SSTO - SSS}{t(r-1)}$ (3.23)

The computation of species mean square and error mean square are crucial for testing the significance of the differences between species means. The null hypothesis tested here is the population means of species are all equal, that is,

$H_0: \mu_1 = \mu_2 = \ldots = \mu_t$

Under this hypothesis, the above two mean squares will be two independent estimates of the same random effect, *i.e.*, *MSS* estimates the same variance as that the *MSE* does. The hypothesis of equal species effects can now be tested by *F* test where *F* is the ratio of *MSS* to *MSE* which follows *F* distribution with (*t*-1) and *t*(*r*-1) degrees of freedom. The significance of *F* can be determined in the usual way by using the table of *F* (Appendix 3). If the calculated value of *F* is larger than the tabled value, the hypothesis is rejected. It implies that at least one pair of species is significantly different with respect to the observations made.

The above results can be presented in the form of a table called analysis of variance table or simply ANOVA table. The format of ANOVA table is as follows.

Table 3.4. ANOVA table

Sources of variation	Degrees of freedom *(df)*	Sum of squares *(SS)*	Mean squares $\left(MS = \frac{SS}{df}\right)$	Computed *F*-ratio
Between species	*t*-1	*SSS*	*MSS*	$\frac{MSS}{MSE}$
Within species (error)	*t*(*r*-1)	*SSE*	*MSE*	
Total	*tr*-1	*SSTO*		

For the purpose of illustration, consider the data presented in Table 3.5. The data represent a set of observations on wood density obtained on a randomly collected set of stems belonging to five cane species.

The analysis of variance for the sample data is conducted as follows.

Step 1. Compute the species totals, species means, grand total and grand mean (as in Table 3.5). Here, the number of species = t = 5 and number of replication = r = 3.

Table 3.5. Wood density (g/cc) observed on a randomly collected set of stems belonging to different cane species.

	Species					
	1	2	3	4	5	Overall
1	0.58	0.53	0.49	0.53	0.57	
2	0.54	0.63	0.55	0.61	0.64	
3	0.38	0.68	0.58	0.53	0.63	
Total	1.50	1.85	1.62	1.67	1.85	8.49
Mean	0.50	0.62	0.54	0.56	0.62	0.57

Step 2. Compute the correction factor *C.F* using Equation (3.18).

$$C.F. = \frac{(8.49)^2}{(5)(3)} = 4.81$$

Step 3. Compute the total sum of squares using Equation (3.21).

$$SSTO = (0.58)^2 + (0.53)^2 + \ldots + (0.63)^2 - \frac{(8.49)^2}{(5)(3)}$$

$$= 0.0765$$

Step 4. Compute the species sum of squares using Equation (3.17).

$$SSS = \frac{(1.50)^2 + (1.84)^2 + \ldots + (1.84)^2}{5} - \frac{(8.49)^2}{(5)(3)}$$

$$= 0.0307$$

Step 5. Compute the error sum of squares as $SSE = SSTO - SSS$

$$SSE = 0.0765 - 0.0307$$

$$= 0.0458$$

Step 6. Compute the mean squares for species and error. These are obtained using Equations (3.22) and (3.23).

$$MSS = \frac{0.0307}{5-1}$$

$$= 0.0153$$

$$MSE = \frac{0.0458}{5(3-1)}$$

$$= 0.0038$$

Step 7. Calculate the F ratio as

$$F = \frac{\text{Treatment } MS}{\text{Error } MS}$$

$$= \frac{0.0153}{0.0038}$$

$$= 4.0108$$

Step 8. Summarise the results as shown in Table 3.6.

Table 3.6. ANOVA table for the data in Table 3.5.

Sources of variation	Degree of freedom *(df)*	Sum of squares *(SS)*	Mean squares $\left(MS = \frac{SS}{df}\right)$	Computed F-ratio	Tabular F

Between species	4	0.0307	0.0153	4.01	3.48
Within species	10	0.0458	0.0038		
Total	14	0.0765			

Compare the computed value of *F* with tabular value of *F* at 4 and 10 degrees of freedom. In this example, the computed value of *F* (1.73) is less than the tabular value (3.48) at 5% level of significance. It may thus be concluded that there are no significant differences among the means of different species.

3.7. Transformation of data

As indicated in the previous section, the validity of analysis of variance depends on certain important assumptions. The analysis is likely to lead to faulty conclusions when some of these assumptions are violated. A very common case of violation is the assumption regarding the constancy of variance of errors. One of the alternatives in such cases is to go for a weighted analysis of variance wherein each observation is weighted by the inverse of its variance. For this, an estimate of the variance of each observation is to be obtained which may not be feasible always. Quite often, the data are subjected to certain scale transformations such that in the transformed scale, the constant variance assumption is realized. Some of such transformations can also correct for departures of observations from normality because unequal variance is many times related to the distribution of the variable also. Certain methods are available for identifying the transformation needed for any particular data set (Montgomery and Peck, 1982) but one may also resort to certain standard forms of transformations depending on the nature of the data. The most common of such transformations are *logarithmic transformation*, *square root transformation* and *angular transformation.*

3.7.1. Logarithmic transformation

When the data are in whole numbers representing counts with a wide range, the variances of observations within each group are usually proportional to the squares of the group means. For data of this nature, logarithmic transformation is recommended. A simple plot of group means against the group standard deviation will show linearity in such cases. A good example is data from an experiment involving various types of insecticides. For the effective insecticide, insect counts on the treated experimental unit may be small while for the ineffective ones, the counts may range from 100 to several thousands. When zeros are present in the data, it is advisable to add 1 to each observation before making the transformation. The log transformation is particularly effective in normalising positively skewed distributions. It is also used to achieve additivity of effects in certain cases.

3.7.2. Square root transformation

If the original observations are brought to square root scale by taking the square root of each observation, it is known as square root transformation. This is appropriate when the variance is proportional to the mean as discernible from a graph of group variances against group means. Linear relationship between mean and variance is commonly observed when the data are in the form of small whole numbers (*e.g.*, counts of wildlings per quadrat, weeds per plot, earthworms per square metre of soil, insects caught in traps, etc.). When the observed values fall within the range of 1 to 10 and especially when zeros are present, the transformation should be, $\sqrt{y+0.5}$. The transformation of the type $\sqrt{y+(3/8)}$ is also used for certain theoretical reasons.

3.7.3. Angular transformation

In the case of proportions, derived from frequency data, the observed proportion p can be changed to a new form $\theta = \sin^{-1}\sqrt{p}$. This type of transformation is known as angular or arcsin transformation. However, when nearly all values in the data lie between 0.3 and 0.7, there is no need for such transformation. It may be noted that the angular transformation is not applicable to proportion or percentage data which are not derived from counts. For example, percentage of marks, percentage of profit, percentage of protein in grains, oil content in seeds, etc., can not be subjected to angular transformation. The angular transformation is not good when the data contain 0 or 1 values for p. The transformation in such cases is improved by replacing 0 with $(1/4n)$ and 1 with $[1-(1/4n)]$, before taking angular values, where n is the number of observations based on which p is estimated for each group.

As a case of illustration of angular transformation, consider the data given in Table 3.7 which represents the percentage of rooting obtained after sixth months of applying hormone treatment at different levels to stem cuttings of a tree species. Three batches, each of 10 stem cuttings, were subjected to dipping treatment in hormone solution at each level. The hormone was tried at 3 levels and the experiment had an untreated control. The percentage of rooting for each batch of cuttings was arrived at by dividing the number of cuttings rooted by the number of cuttings in a batch.

Table 3.7. Percentage of rooting obtained at the sixth month after applying the treatments.

	Treatments			
Batch of cuttings	Control	IBA at 10 ppm	IBA at 50 ppm	IBA at 100 ppm
1	0	70	60	30
2	0	80	70	20
3	0	60	70	10

The data in Table 3.7 was transformed to angular scale using the function, $\sin^{-1}\sqrt{p}$ after replacing the '0' values with $(1/4n)$ where n =10. The functional values of

$\sin^{-1}\sqrt{p}$ for different values of p can also be obtained from Table (X) of Fisher and Yates (1963). The transformed data of Table 3.7 is given in Table 3.8.

Table 3.8. The data of Table 3.7 transformed to angular scale.

Batch of cuttings	Treatments				Grand total
	Control	IBA at 10 ppm	IBA at 50 ppm	IBA at 100 ppm	
1	0.99	56.79	50.77	33.21	
2	0.99	63.44	56.79	26.56	
3	0.99	50.77	56.79	18.44	
Total	2.97	171	164.35	78.21	416.53

In order to see if the treatments differ significantly in their effects, a one-way analysis of variance can be carried out as described in Section 3.6, on the transformed data. The results of analysis of variance are presented in Table 3.9.

Table 3.9. Analysis of variance of the transformed data in Table 3.8.

Sources of variation	Degrees of freedom *(df)*	Sum of squares *(SS)*	Mean square $\left(MS = \frac{SS}{df}\right)$	Computed *F*-ratio	Tabular *F* at 5% level
Between treatments	3	6334.41	2111.47	78.96*	4.07
Within treatments	8	213.93	26.74		
Total	11	6548.34			

* significant at 5% level.

Before concluding this section, a general note is added here. Once the transformation has been made, the analysis is carried out with the transformed data and all the conclusions are drawn in the transformed scale. However, while presenting the results, the means and their standard errors are transformed back into original units. While transforming back into the original units, certain corrections have to be made for the means. In the case of log transformed data, if the mean value is $\bar{y}$, the mean value of the original units will be $\text{antilog}(\bar{y} + 1.15\bar{y})$ instead of $\text{antilog}(\bar{y})$. If the square root transformation had been used, then the mean in the original scale would be $\left(\bar{y} + V(\bar{y})\right)^2$ instead of $\left(\bar{y}\right)^2$ where $V(\bar{y})$ represents the variance of $\bar{y}$. No such correction is generally made in the case of angular transformation. The inverse transformation for angular transformation would be $p = (\sin\theta)^2$.

3.8. Correlation

In many natural systems, changes in one attribute are accompanied by changes in another attribute and that a definite relation exists between the two. In other words, there is a correlation between the two variables. For instance, several soil properties

like nitrogen content, organic carbon content or pH are correlated and exhibit simultaneous variation. Strong correlation is found to occur between several morphometric features of a tree. In such instances, an investigator may be interested in measuring the strength of the relationship. Having made a set of paired observations (x_i,y_i); $i = 1, ..., n$, from n independent sampling units, a measure of the linear relationship between two variables can be obtained by the following quantity called Pearson's product moment correlation coefficient or simply correlation coefficient.

$$r = \frac{\text{Covariance of } x \text{ and } y}{\sqrt{(\text{Variance of } x)(\text{Variance of } y)}} = \frac{Cov(x,y)}{\sqrt{(V(x))(V(y))}} \qquad (3.24)$$

where $Cov\ (x,y) = \frac{1}{n}\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right)$

$$V(x) = \frac{1}{n}\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)$$

$$V(y) = \frac{1}{n}\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right) = \frac{1}{n}\left(\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right)$$

It is a statistical measure which indicates both the direction and degree of relationship between two measurable characteristics, say, x and y. The range of r is from -1 to +1 and does not carry any unit. When its value is zero, it means that there is no linear relationship between the variables concerned (although it is possible that a nonlinear relationship exists). A strong linear relationship exists when the value of r approaches -1 or +1. A negative value of r is an indication that an increase in the value of one variable is associated with a decrease in the value of other. A positive value on the other hand, indicates a direct relationship, *i.e.*, an increase in the value of one variable is associated with an increase in the value of the other. The correlation coefficient is not affected by change of origin or scale or both. When a constant term is added or subtracted from the values of a variable, we say that the origin is changed. Multiplying or dividing the values of a variable by a constant term amounts change of scale.

As an example, consider the data on pH and organic carbon content measured from soil samples collected from 15 pits taken in natural forests, given in Table 3.10.

Table 3.10. Values of pH and organic carbon content observed in soil samples collected from natural forest.

Soil pit	pH (x)	Organic carbon (%) (y)	(x^2)	(y^2)	(xy)
1	5.7	2.10	32.49	4.4100	11.97
2	6.1	2.17	37.21	4.7089	13.24
3	5.2	1.97	27.04	3.8809	10.24
4	5.7	1.39	32.49	1.9321	7.92
5	5.6	2.26	31.36	5.1076	12.66
6	5.1	1.29	26.01	1.6641	6.58
7	5.8	1.17	33.64	1.3689	6.79
8	5.5	1.14	30.25	1.2996	6.27
9	5.4	2.09	29.16	4.3681	11.29
10	5.9	1.01	34.81	1.0201	5.96
11	5.3	0.89	28.09	0.7921	4.72
12	5.4	1.60	29.16	2.5600	8.64
13	5.1	0.90	26.01	0.8100	4.59
14	5.1	1.01	26.01	1.0201	5.15
15	5.2	1.21	27.04	1.4641	6.29
Total	82.1	22.2	450.77	36.4100	122.30

The steps to be followed in the computation of correlation coefficient are as follows.

Step 1. Compute covariance of x and y and variances of both x and y using Equation (3.24).

$$Cov\,(x,y) = \frac{1}{15}\left(122.30 - \frac{(82.1)(22.2)}{15}\right)$$

$$= 0.05$$

$$V(x) = \frac{1}{15}\left(450.77 - \frac{(82.1)^2}{15}\right)$$

$$= 0.0940$$

$$V(y) = \frac{1}{15}\left(36.41 - \frac{(22.2)^2}{15}\right)$$

$$= 0.2367$$

Step 2. Compute the correlation coefficient using Equation (3.24).

$$r = \frac{0.05}{\sqrt{(0.0940)(0.2367)}}$$

$$= 0.3541$$

3.8.1. Testing the significance of correlation coefficient.

A value of correlation coefficient obtained from a sample needs to be tested for significance to confirm if a real relationship exists between the two variables in the population considered. It is usual to set up the null hypothesis as $H_0: \rho = 0$ against the alternative hypothesis, $H_1: \rho \neq 0$.

For relatively small *n*, the null hypothesis that $\rho = 0$ can be tested using the test statistic,

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad (3.25)$$

This test statistic is distributed as Student's *t* with *n*-2 degrees of freedom.

Consider the data given in Table 3.10 for which n = 15 and r = 0.3541. To test as $H_0: \rho = 0$ against $H_1: \rho \neq 0$, we compute the test statistic using Equation (3.25).

$$t = \frac{0.3541\sqrt{15-2}}{\sqrt{1-(0.3541)^2}} = 1.3652$$

Referring Appendix 2, the critical value of t is 2.160, for 13 degrees of freedom at the probability level, α = 0.05. Since the computed *t* value is less than the critical value, we conclude that the pH and organic carbon content measured from soil samples are not significantly correlated. For convenience one may use Appendix 5 which gives values of correlation coefficients beyond which an observed correlation coefficient can be declared as significant for a certain number of observations at a desired level of significance.

In order to test the hypothesis that $H_0: \rho = \rho_0$ where ρ_0 is any specified value of ρ, Fisher's *z*-transformation is employed which is given by,

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \qquad (3.26)$$

where ln indicates natural logarithm.

For testing the null hypothesis, we use the test statistic,

$$w = \frac{z - z_0}{\sqrt{\frac{1}{n-3}}} \qquad (3.27)$$

where $z_0 = \frac{1}{2}\ln\left(\frac{1+\rho_o}{1-\rho_0}\right)$

The statistic *w* follows a standard normal distribution.

For the purpose of illustration, consider the data given in Table 3.10 for which n = 15 and r = 0.3541. Suppose that we have to test $H_0: \rho = \rho_0 = 0.6$. For testing this null hypothesis, the values of r and ρ are first subjected to z transformation.

$$z = \frac{1}{2}\ln\left(\frac{1+0.3541}{1-0.3541}\right) = 0.3701$$

$$z_0 = \frac{1}{2}\ln\left(\frac{1+0.6}{1-0.6}\right) = 0.6932$$

The value of the test statistic would be,

$$w = \frac{0.3701 - 0.6932}{\sqrt{\frac{1}{15-3}}} = 1.16495$$

Since the value of *w* is less than 1.96, the critical value, it is nonsignificant at 5% level of significance. Hence we may conclude that the correlation coefficient between pH and organic carbon content in the population is not significantly different from 0.6.

3.9. Regression

Correlation coefficient measures the extent of interrelation between two variables which are simultaneously changing with mutually extended effects. In certain cases, changes in one variable are brought about by changes in a related variable but there need not be any mutual dependence. In other words, one variable is considered to be dependent on the other variable changes, in which are governed by extraneous factors. Relationship between variables of this kind is known as regression. When such relationships are expressed mathematically, it will enable us to predict the value of one variable from the knowledge of the other. For instance, the photosynthetic and transpiration rates of trees are found to depend on atmospheric conditions like temperature or humidity but it is unusual to expect a reverse relationship. However, in many cases, it so happens that the declaration of certain variables as independent is made only in a statistical sense although when reverse effects are conceivable in such cases. For instance, in a volume prediction equation, tree volume is taken to be dependent on dbh although the dbh cannot be considered as independent of the effects of tree volume in a physical sense. For this reason, independent variables in the context of regression are sometimes referred as regressor variables and the dependent variable is called the regressand.

The dependent variable is usually denoted by y and the independent variable by x. When only two variables are involved in regression, the functional relationship is known as *simple regression*. If the relationship between the two variables is linear, it is known as *simple linear regression*, otherwise it is known as *nonlinear regression*. When one variable is dependent on two or more independent variables, the functional relationship between the dependent and the set of independent variables is known as *multiple regression*. For the sake of easiness in description, only the case of simple linear regression is considered here. Reference is made to Montgomery and Peck (1982) for more complex cases.

3.9.1. Simple linear regression

The simple linear regression of y on x in the population is expressible as

$$y = \alpha + \beta x + \varepsilon \qquad (3.28)$$

where α and β are parameters also known as regression coefficients and ε is a random deviation possible from the expected relation. But for ε with a mean value of zero, Equation (3.28) represents a straight line with α as the intercept and β as the slope of the line. In other words, α is the expected value of y when x assumes the value zero and β gives the expected change in y for a unit change in x. The slope of a linear regression line may be positive, negative or zero depending on the relation between y and x.

In practical applications, the values of α and β are to be estimated from observations made on y and x variables from a sample. For instance, to estimate the parameters of a regression equation proposed between atmospheric temperature and transpiration rate of trees, a number of paired observations are made on transpiration rate and temperature at different times of the day from a number of trees. Let such pairs of values be designated as (x_i, y_i); $i = 1, 2, \ldots, n$ where n is the number of independent pairs of observations. The values of α and β are estimated using the method of least squares (Montgomery and Peck, 1982) such that the sum of squares of the difference between the observed and expected value is minimum. In the estimation process, the following assumptions are made *viz*., (i) The x values are non-random or fixed (ii) For any given x, the variance of y is the same (iii) The y values observed at different values of x are completely independent. Appropriate changes will need to be made in the analysis when some of these assumptions are not met by the data. For the purpose of testing hypothesis of parameters, an additional assumption of normality of errors will be required.

In effect, the values of α and β are obtained from the formulae,

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i-1}^{n} x_i\right)^2}{n}} \quad (3.29)$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \quad (3.30)$$

The equation $\hat{y} = \hat{\alpha} + \hat{\beta}x$ is taken as the fitted regression line which can be used to predict the average value of the dependent variable, *y*, associated with a particular value of the independent variable, *x*. Generally, it is safer to restrict such predictions within the range of *x* values in the data.

The standard errors of $\hat{\beta}$ and $\hat{\alpha}$ can be estimated by the following formulae.

$$SE(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i-1}^{n} x_i\right)^2}{n}}} \quad (3.31)$$

$$SE(\hat{\alpha}) = \sqrt{\frac{\hat{\sigma}^2 \frac{\sum_{i=1}^{n} x_i^2}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i-1}^{n} x_i\right)^2}{n}}} \quad (3.32)$$

where $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{n}$

The standard error of an estimate is the standard deviation of the sampling distribution of that estimate and is indicative of the extent of reliability of that estimate.

As an example, consider the data presented in Table 3.11 which contain paired values of photosynthetic rate and light interception observed on leaves of a particular tree species. In this example, the dependent variable is photosynthetic rate and the independent variable is the quantity of light. The computations involved in fitting a regression line are given in the following.

Step 1. Compute the values of numerator and denominator of Equation (3.29) using the sums, sum of squares and sum of products of *x* and *y* generated in Table 3.5.

$$\sum xy - \frac{\sum x \sum y}{n} = 175.59 - \frac{(13.72)(189.03)}{15} = 2.6906$$

$$\sum x^2 - \frac{\left(\sum x\right)^2}{n} = 12.70 - \frac{(13.72)^2}{15} = 0.1508$$

Table 3.11. Data on photosynthetic rate in μ mol $m^{-2}s^{-1}$ (y) along with the measurement of radiation in mol $m^{-2}s^{-1}$ (x) observed on a tree species.

X	y	x^2	xy
0.7619	7.58	0.58	5.78
0.7684	9.46	0.59	7.27
0.7961	10.76	0.63	8.57
0.8380	11.51	0.70	9.65
0.8381	11.68	0.70	9.79
0.8435	12.68	0.71	10.70
0.8599	12.76	0.74	10.97
0.9209	13.73	0.85	12.64
0.9993	13.89	1.00	13.88
1.0041	13.97	1.01	14.02
1.0089	14.05	1.02	14.17
1.0137	14.13	1.03	14.32
1.0184	14.20	1.04	14.47
1.0232	14.28	1.05	14.62
1.0280	14.36	1.06	14.77
$\sum x = 13.72$	$\sum y = 189.03$	$\sum x^2 = 12.70$	$\sum xy = 175.59$

Step 2. Compute the estimate of α and β using Equations (3.29) and (3.30).

$$\hat{\beta} = \frac{2.6906}{0.1508} = 17.8422$$
$$\hat{\alpha} = 12.60 - (17.8421)(0.9148)$$
$$= -3.7202$$

The fitted regression line is $\hat{y} = -3.7202 + 17.8422x$ which can be used to predict the value of photosynthetic rate at any particular level of radiation within the range of data. Thus, the expected photosynthetic rate at 1 mol $m^{-2}s^{-1}$ of light would be,

$$\hat{y} = -3.7202 + 17.8422(1) = 14.122$$

Step 3. Get an estimate of σ^2 as defined in Equation (3.32).

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}\right)^2}{n} = 0.6966$$

Step 4. Develop estimates of standard errors of $\hat{\beta}$ and $\hat{\alpha}$ using Equations (3.31) and (3.32).

$$SE(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}} = \sqrt{\frac{0.6966}{12.70 - \frac{(13.72)^2}{15}}} = 2.1495$$

$$SE(\hat{\alpha}) = \sqrt{\frac{\hat{\sigma}^2 \frac{\sum x^2}{n}}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}} = \sqrt{\frac{0.6966 \frac{12.70}{15}}{12.70 - \frac{(13.72)^2}{15}}} = 1.9778$$

3.9.2. Testing the significance of the regression coefficient

Once the regression function parameters have been estimated, the next step in regression analysis is to test the statistical significance of the regression function. It is usual to set the null hypothesis as $H_0: \beta = 0$ against the alternative hypothesis, $H_1: \beta \neq 0$ or ($H_1: \beta < 0$ or $H_1: \beta > 0$, depending on the anticipated nature of relation). For the testing, we may use the analysis of variance procedure. The concept of analysis of variance has already been explained in Section 3.6 but its application in the context of regression is shown below using the data given in Table 3.11.

Step1. Construct an outline of analysis of variance table as follows.

Table 3.12. Schematic representation of analysis of variance for regression analysis.

Source of variation	Degree of freedom (*df*)	Sum of squares (*SS*)	Mean square $\left(MS = \frac{SS}{df}\right)$	Computed *F*
Due to regression	1	*SSR*	*MSR*	$\frac{MSR}{MSE}$
Deviation from regression	*n*-2	*SSE*	*MSE*	
Total	*n*-1	*SSTO*		

Step 2. Compute the different sums of squares as follows.

Total sum of squares = $SSTO = \sum y^2 - \frac{\left(\sum y\right)^2}{n}$ (3.33)

$$= (7.58)^2 + (9.46)^2 + \ldots + (14.36)^2 - \frac{(189.03)^2}{15}$$

$$= 58.3514$$

$$\text{Sum of square due to regression} = SSR = \frac{\left[\sum xy - \frac{\sum x \sum y}{n}\right]^2}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}} \quad (3.34)$$

$$= \frac{(2.6906)^2}{0.1508}$$

$$= 48.0062$$

Sum of squares due to deviation from regression = $SSE = SSTO - SSR$ (3.35)

$$= 58.3514 - 48.0062 = 10.3452$$

Step 3. Enter the values of sums of squares in the analysis of variance table as in Table 3.13 and perform the rest of the calculations.

Table 3.13. Analysis of variance for the regression equation derived for data in Table 3.11.

Source of variation	Degree of freedom (*df*)	Sum of squares (*SS*)	Mean square $\left(MS = \frac{SS}{df}\right)$	Computed *F* at 5%
Due to regression	1	48.0062	48.0062	60.3244
Deviation from regression	13	10.3452	0.7958	
Total	14	58.3514		

Step 4. Compare the computed value of *F* with tabular value at (1,*n*-2) degrees of freedom. For our example, the calculated value (60.3244) is greater than the tabular value of *F* of 4.67 at (1,13) degrees of freedom at 5% level of significance and so the *F* value is significant. If the computed *F* value is significant, then we can state that the regression coefficient, β is significantly different from 0. The sum of squares due to regression when expressed as a proportion of the total sum of squares is known as the coefficient of determination which measures the amount of variation in *y* accounted by the variation in *x*. In other words, coefficient of determination indicates the proportion of the variation in the dependent variable explained by the model. For the present example, the coefficient of determination (R^2) is

$$R^2 = \frac{SSR}{SSTO} \quad (3.36)$$

$$= \frac{48.0062}{58.3514}$$

$$= 0.8255$$

3.10. Analysis of covariance

In analysis of variance, generally, the significance of any known component of variation is assessed in comparison to the unexplained residual variation. Hence, proper control is to be exercised to reduce the magnitude of the uncontrolled variation. Either the model is expanded by including more known sources of variation or deliberate control is made on many variables affecting the response. Otherwise, any genuine group differences would go undetected in the presence of the large residual variation. In many instances, the initial variation existing among the observational units is largely responsible for the variation in their further responses and it becomes necessary to eliminate the influence of inherent variation among the subjects from the comparison of the groups under consideration. Analysis of covariance is one of the methods used for the purpose of reducing the magnitude of unexplained error. For instance, in an experimental context, covariance analysis can be applied when observations on one or more correlated variables are available from each of the experimental units along with the observations on the response variable under study. These additional related variables are called covariates or ancillary or concomitant variables. It is necessary that these variables are associated with the variable under study. For example, in yield trials, variation in the initial stocking induced by extraneous factors, residual effects of the previous crops grown in the site etc., can serve as covariates.

Analysis of covariance is a synthesis of the methods of the analysis of variance and those of regression. The concept is elaborated here in the context of an experiment with just one variable under study denoted by y and a single covariate denoted by x. Let there be t experimental groups to be compared, each group consisting of r experimental units. The underlying model in this case could be

$$y_{ij} = \mu_y + \alpha_i + \beta(x_{ij} - \mu_x) + e_{ij} \qquad (3.37)$$

where y_{ij} is the response observed on the jth experimental unit belonging to ith group, ($i = 1, 2, \ldots, t$; $j = 1, 2, \ldots, r$)

μ_y is the overall population mean of y,

α_i is the effect of being in the ith group,

β is the within group regression coefficient of y on x

x_{ij} is the observation on ancillary variate on jth unit of ith group.

μ_x is the overall mean of the covariate

e_{ij} 's are the error components which are assumed to be normally and independently distributed with zero mean and a constant variance s^2.

Covariance analysis is essentially an extension of the analysis of variance and hence, all the assumptions for a valid analysis of variance apply here as well. In addition, the covariance analysis requires that (i) The relationship between the primary character of interest y and the covariate x is linear (ii) The strength of the relation between y and x

remains the same in each experimental group (iii) The variation in the covariate should not have been a result of the group differences.

The steps involved in the analysis of covariance are explained below.

Step 1. The first step in the analysis of covariance is to compute the sum of squares due to the different components, for the variate y and the covariate x in the usual manner of analysis of variance. The computation formulae for the same are given below.

Total SS of y = $SSTO(y)$ = $\sum_{i=1}^{t}\sum_{j=1}^{r} y_{ij}^2 - C.F.(y)$ (3.38)

where $C.F.(y) = \dfrac{\left(\sum_{i=1}^{t}\sum_{j=1}^{r} y_{ij}\right)^2}{tr}$ (3.39)

Group SS of y = $SSG(y)$ = $\dfrac{\sum_{i=1}^{t} y_{i.}^2}{r} - C.F.(y)$ (3.40)

Error SS of y = $SSE(y)$ = $SSTO(y)$ -$SSG(y)$ (3.41)

Total SS of x = $SSTO(x)$ = $\sum_{i=1}^{t}\sum_{j=1}^{r} x_{ij}^2 - C.F.(x)$ (3.42)

where $C.F.(x) = \dfrac{\left(\sum_{i=1}^{t}\sum_{j=1}^{r} x_{ij}\right)^2}{tr}$ (3.43)

Group SS of x = $SSG(x)$ = $\dfrac{\sum_{i=1}^{t} x_{i.}^2}{r} - C.F.(x)$ (3.44)

Error SS of x = $SSE(x)$ = $SSTO(x)$ - $SSG(x)$ (3.45)

Step 2. Compute the sum of products for x and y as follows.

Total SP = $SPTO(xy)$ = $\sum_{i=1}^{t}\sum_{j=1}^{r} y_{ij}x_{ij} - C.F.(xy)$ (3.46)

where $C.F.(xy) = \dfrac{\left(\sum_{i=1}^{t}\sum_{j=1}^{r} y_{ij}\right)\left(\sum_{i=1}^{t}\sum_{j=1}^{r} x_{ij}\right)}{tr}$ (3.47)

Group SP = $SPG(xy)$ = $\dfrac{\sum_{i=1}^{t} y_{i.}x_{i.}}{r} - C.F.(xy)$ (3.48)

Error SP = $SPE(xy)$ = $SSTO(xy)$ - $SSG(xy)$ (3.49)

Step 3. The next step is to verify whether the covariate is affected by the experimental groups. If x is not affected by the groups, there should not be significant differences between groups with respect to x. The regression co-efficient within groups is computed as

$$\hat{\beta} = \frac{SPE(xy)}{SSE(x)} \tag{3.50}$$

The significance of $\hat{\beta}$ is tested using F-test. The test statistic F is given by

$$F = \frac{\frac{\left(SPE(xy)\right)^2}{SSE(x)}}{\left\{SSE(y) - \frac{\left(SPE(xy)\right)^2}{SSE(x)}\right\} \Big/ \left(t(r-1)-1\right)} \tag{3.51}$$

The F statistic follows a F distribution with 1 and $t(r$-1)-1 degrees of freedom. If the regression coefficient is significant, we proceed to make adjustments in the sum of squares for y for the variation in x. If not significant, it is not worthwhile to make the adjustments.

Step 4. Adjusted values for y are computed as follows:

Adjusted total *SS* of y = Adj. $SSTO(y) = SSTO(y)$ - $\frac{\left(SPTO(xy)\right)^2}{SSTO(y)}$ (3.52)

Adjusted error *SS* of y = Adj. $SSE(y) = SSE(y)$ - $\frac{\left(SPE(xy)\right)^2}{SSE(x)}$ (3.53)

Adjusted group *SS* of y = Adj. $SSG(y)$= Adj. $SSTO(y)$ - Adj. $SSE(y)$ (3.54)

Conventionally, the above results are combined in a single table as in Table 3.14.

Step 5. The adjusted group means are obtained by the formula,

$$\bar{y}_i' = \bar{y}_i - \hat{\beta}\left(\bar{x}_i - \bar{x}\right) \tag{3.55}$$

The standard error for the difference between two adjusted means is given by :

$$SE(d) = \sqrt{MSE\left[\frac{1}{r_i} + \frac{1}{r_j} + \frac{\left(\bar{x}_i - \bar{x}_j\right)^2}{SSE(x)}\right]} \tag{3.56}$$

where the symbols have the usual meanings.

When the number of replications is the same for all the groups and when averaged over all values of $\left(\bar{x}_i - \bar{x}_j\right)^2$ we get,

$$SE(d) = \sqrt{\frac{2\,MSE}{r}\left[1+\frac{SSG(x)}{(t-1)SSE(x)}\right]} \qquad (3.57)$$

Table 3.14. Analysis of covariance (ANOCOVA) table

Source of variation	*df*	Sum of squares and products		
		y	*x*	*xy*
Total	*tr*-1	*SSTO(y)*	*SSTO(x)*	*SPTO(xy)*
Group	*t*-1	*SSG(y)*	*SSG(x)*	*SPG(xy)*
Error	*t*(*r*-1)	*SSE(y)*	*SSE(x)*	*SPE(xy)*

Table 3.14. Cond…

	Adjusted values for *y*			
Source of variation	*df*	*SS*	*MS*	*F*
Total	*tr*-2	Adj. *SSTO(y)*	-	-
Group	-	-	-	-
Error	*t*(*r*-1)-1	Adj. *SSE(y)*	*MSE*	-
Adj. Group	*t*-1	Adj. *SSG(y)*	*MSG*	$\frac{MSG}{MSE}$

Let us use the data given in Table 3.15 to demonstrate how the analysis of covariance is carried out. The data represent plot averages based on forty observations of initial height (*x*) and height attained after four months (*y*) of three varieties of ipil ipil, (*Leucaena leucocephala)*, each grown in 10 plots in an experimental site.

Table 3.15. Initial height (*x*) and height after four months (*y*) in cm of three varieties of ipil ipil (*Leucaena leucocephala)*, in an experimental area.

Plot	Variety 1		Variety 2		Variety 3	
	x	*y*	*x*	*y*	*x*	*y*
1	18	145	27	161	31	180
2	22	149	28	164	27	158
3	26	156	27	172	34	183
4	19	151	25	160	32	175
5	15	143	21	166	35	195
6	25	152	30	175	36	196
7	16	144	21	156	35	187
8	28	154	30	175	23	137
9	23	150	22	158	34	184
10	24	151	25	165	32	184
Total	216	1495	256	1652	319	1789
Mean	21.6	149.5	25.6	165.2	31.2	178.9

The analysis is carried out following the format shown in Table 3.14. The computations are demonstrated below:

Step 1.Compute sum of squares for *x* and *y* variables using Equations (3.38) to (3.45).

$$C.F.(y) = \frac{(4936)^2}{(3)(10)} = 812136.5333$$

$$SSTO(y) = (145)^2 + (149)^2 + \ldots + (184)^2 - 812136.5333$$
$$= 7493.4667$$

$$SSG(y) = \frac{(1495)^2 + (1652)^2 + (1789)^2}{10} - 812136.5333$$
$$= 4328.4667$$

$$SSE(y) = 7493.4667 - 4328.4667$$
$$= 3615.0$$

$$C.F.(x) = \frac{(791)^2}{(3)(10)}$$
$$= 20856.0333$$

$$SSTO(x) = (18)^2 + (22)^2 + \ldots + (32)^2 - 20.856.0333$$
$$= 966.9697$$

$$SSG(x) = \frac{(216)^2 + (256)^2 + (319)^2}{10} - 20856.0333$$
$$= 539.267$$

$$SSE(x) = 966.9697-539.267$$
$$= 427.7027$$

Step 2.Compute sum of products for x and y variables using Equations (3.46) to (3.49).

$$C.F.(xy) = \frac{(791)(4936)}{(3)(10)}$$
$$= 130145.8667$$

$$SPTO(xy) = 18(145) + 22(149) + \ldots + 32(184) - 130145.8667$$
$$= 2407.1333$$

$$SPG(xy) = \frac{216(1495) + 256(1652) + 319(1789)}{10} - 130145.8667$$
$$= 1506.44$$

$$SPE(xy) = 2407.1333 - 1506.44 = 900.6933$$

Step 3. Compute the regression coefficient and test its significance using Equations (3.50) and (3.51).

$$\hat{\beta} = \frac{900.6933}{427.7027}$$
$$= 2.1059$$

The significance of $\hat{\beta}$ is tested using *F*-test. The test statistic *F* is given by the Equation (3.51).

$$F = \frac{\frac{(900.6933)^2}{427.7027}}{\left\{3615 - \frac{(900.6933)^2}{427.7027}\right\} \Big/ (3(10-1)-1)}$$

$$= \frac{1896.7578}{66.0862}$$
$$= 28.7012$$

Table value of *F* with (1,26) degrees of freedom = 9.41 at 5% level of significance. Here calculated value of *F* is grater than tabular value and hence β is significantly different from zero.

Step 4. Compute adjusted sums of squares for the different sources in the ANOCOVA using Equations (3.52) to (3.54). Summarise the results as in Table 3.14 and compute mean square values for group (*MSG*) and error (*MSE*) and also the value of *F* based on these mean squares.

$$\text{Adj. } SSTO(y) = 7493.4667 - \frac{2407.1333^2}{966.9697}$$
$$= 1501.2513$$
$$\text{Adj. } SSE(y) = 3165 - \frac{900.6933}{427.7027}$$
$$= 1268.2422$$

$$\text{Adj. } SSG(y) = 1501.2513 - 1268.2422$$
$$= 233.0091$$

$$MSG = \frac{233.0091}{2} = 116.5046$$
$$MSE = \frac{1268.2422}{3(10-1)-1}$$
$$= 48.7785$$

$$F = \frac{MSG}{MSE}$$

$$= \frac{116.5046}{48.7785}$$
$$= 2.39$$

Table 3.16. Analysis of covariance table for the data in Table 3.15.

Sources of variation	*df*	Sum of squares and products			Adjusted values for *y*			
		y	*x*	*xy*	*df*	*SS*	*MS*	*F*
Total	29	7493.467	966.970	2407.133	28	1501.25	-	-
Group	2	4328.467	539.267	1506.440	-	-	-	-
Error	27	3615.000	427.703	900.693	26	1268.24	48.8	-
Group adjusted for the covariate					2	233.009	116.5	2.4

The value of *F* for (2,26) degrees of freedom at 5% level of significance is 3.37. Since the observed *F* value, 2.4, is less than the critical value, we conclude that there are no significant differences among the varieties.

Step 5. Get the adjusted group means and standard error of the difference between any two adjusted group means by using Equations (3.55) and (3.57).

$$\bar{y}_1' = \bar{y}_1 - \hat{\beta}(\bar{x}_1 - \bar{x}) = 149.5 - 2.1059(21.6 - 26.37) = 159.54$$
$$\bar{y}_2' = \bar{y}_2 - \hat{\beta}(\bar{x}_2 - \bar{x}) = 165.2 - 2.1059(25.6 - 26.37) = 166.82$$
$$\bar{y}_3' = \bar{y}_3 - \hat{\beta}(\bar{x}_3 - \bar{x}) = 178.9 - 2.1059(31.2 - 26.37) = 168.73$$

$$SE(d) = \sqrt{\frac{2\,MSE}{r}\left[1 + \frac{SSG(x)}{(t-1)SSE(x)}\right]}$$
$$= \sqrt{\frac{(2)(48.8)}{10}\left[1 + \frac{539.267}{(3-1)(427.703)}\right]} = 3.9891$$

The standard error of the difference between group means will be useful in pairwise comparison of group means as explained in Chapter 4.

3.11 Analysis of repeated measures

Repeated measurements of observational units are very frequent in forestry research. The term repeated is used to describe measurements which are made of the same characteristic on the same observational unit but on more than one occasion. In longitudinal studies, individuals may be monitored over a period of time to record the changes occurring in their states. Typical examples are periodical measurements on

diameter or height of trees in a silvicultural trial, observations on disease progress on a set of seedlings in a nursery trial, etc. Repeated measures may be spatial rather than temporal. For instance, consider measurements on wood characteristics of several stems at the bottom, middle and top portion of each stem and each set of stems coming from a different species. Another example would be that of soil properties observed from multiple core samples at 0-15, 15-50 and 50-100 cm depth from different types of vegetation.

The distinctive feature of repeated measurements is the possibility of correlations between successive measurements over space or time. Autocorrelation among the residuals arising on account of repeated measurements on the same experimental units leads to violation of the basic assumption of independence of errors for conducting an ordinary analysis of variance. However, several different ways of analysing repeated measurements are available. These methods vary in their efficiency and appropriateness depending upon the nature of data. If the variance of errors at each of the successive measurements is the same and also the covariances between errors of different measurement occasions are the same, one may choose to subject the data to a 'univariate mixed model analysis'. In case the errors are unstructured, a multivariate analysis is suggestible taking repeated measurements as different characters observed on the same entities (Crowder and Hand, 1990). The details of univariate analysis are illustrated below under a simplified observational set up whereas the reader is referred to (Crowder and Hand, 1990) for multivariate analysis in this context.

The general layout here is that of n individuals x p occasions with the individuals divided into t groups of size n_i ($i = 1, 2, \ldots, t$). Let the hypothesis to be tested involve a comparison among the groups. The model used is

$$y_{ijk} = \mu + \alpha_i + e_{ij} + \beta_j + \gamma_{ij} + e_{ijk} \qquad (3.58)$$

where y_{ijk} is the observation on kth individual in the ith group at jth occasion;
($i = 1, \ldots, t, j = 1, \ldots, p, k = 1, \ldots, n_i$.)

μ is the general mean,

α_i is the effect of ith level of the factor 'group',

β_j is the effect of jth level of the factor 'occasion',

γ_{ij} is the interaction effect for the ith level of the factor 'group' and jth level of the factor 'occasion'. This term measures the differences between the groups with respect to their pattern of response over occasions. More elaborate discussion on interaction is included in Chapter 4.

In model (3.58), the random component e_{ij} are assumed to be independently and normally distributed with mean zero and variance s_e^2 and e_{ijk} is the random error component which is also assumed to be independently and normally distributed with mean zero and variance s_w^2. In the model, α_i's and β_j's are assumed to be fixed.

Let $y_{i..}$ denote the total of all observations under the ith level of factor, group; $y_{.j.}$ denote the total of all observations under the jth level of factor occasion; $y_{ij.}$ denote the total of all observations in the (ij)th cell; $y_{...}$ denote the grand total of all the observations. These notations are expressed mathematically as

$$y_{i..} = \sum_{j}^{p}\sum_{k}^{n_i} y_{ijk}\ ,\ y_{.j.} = \sum_{i}^{t}\sum_{k}^{n_i} y_{ijk}\ ,\ y_{ij.} = \sum_{k}^{n_i} y_{ijk}\ ,\ y_{...} = \sum_{i}^{t}\sum_{j}^{p}\sum_{k}^{n_i} y_{ijk}$$

The univariate mixed model ANOVA is shown below.

Table 3.17. Schematic representation of univariate mixed model analysis of variance.

Sources of variation	Degrees of freedom	Sum of squares	Mean sum of squares	*F*-ratio
Group	t-1	SSG	$MSG = \frac{SSG}{t-1}$	$\frac{MSG}{MSE_a}$
Individuals within groups	$\sum_i (n_i - 1)$	SSE_a	$MSE_a = \frac{SSE_a}{\sum_i (n_i - 1)}$	
Occasion	p-1	SSO	$MSO = \frac{SSO}{p-1}$	$\frac{MSO}{MSE_b}$
Occasion x Group	(t-1)(p-1)	$SSOG$	$MSOG = \frac{SSOG}{(t-1)(p-1)}$	$\frac{MSOG}{MSE_b}$
Occasion x Individuals within groups	$(p-1)\sum_i (n_i - 1)$	SSE_b	$MSE_b = \frac{SSE_b}{(p-1)\sum_i (n_i - 1)}$	
Total	$p\sum_i n_i - 1$	$SSTO$		

The computational formulae for the sum of squares in the above table are as follows,

$$SSTO = \sum_i \sum_j \sum_k y_{ijk}^2 - \frac{y_{...}^2}{p\sum_i n_i} \qquad (3.59)$$

$$SSG = \sum_i \frac{y_{i..}^2}{pn_i} - \frac{y_{...}^2}{p\sum_i n_i} \qquad (3.60)$$

$$SSE_a = \sum_i \sum_k \frac{y_{i.k}^2}{p} - \sum_i \frac{y_{i..}^2}{pn_i} \qquad (3.61)$$

$$SSO = \sum_j \frac{y_{.j.}^2}{\sum_i n_i} - \frac{y_{...}^2}{p\sum_i n_i} \quad (3.62)$$

$$SSOG = \sum_i \sum_j \frac{y_{ij}^2}{n_i} - \sum_i \frac{y_{i...}^2}{pn_i} - \sum_j \frac{y_{.j,}^2}{\sum_i n_i} + \frac{y_{...}^2}{p\sum_i n_i} \quad (3.63)$$

$$SSE_b = SST - SSG - SSE_a - SSO - SSOG \quad (3.64)$$

For illustration of the analysis, consider the data given in Table 3.18. The data represent the mycelial growth (mm) of five isolates of *Rizoctonia solani* on PDA medium after 14, 22, 30 and 38 hours of incubation, each isolate grown in three units of the medium. Here, the isolates represent 'groups' and different time points represent the 'occasions' of Table 3.17.

Table 3.18. Data on mycelial growth (mm) of five groups of *R. solani* isolates on PDA medium.

		Mycelial growth (mm) observed at different occasions			
R. Solani isolate	PDA unit	14 hr.	22 hr.	30 hr.	38 hr.
1	1	29.00	41.00	55.00	68.50
	2	28.00	40.00	54.00	68.50
	3	29.00	42.00	55.00	69.00
2	1	33.50	46.50	59.00	74.00
	2	31.50	44.50	58.00	71.50
	3	29.00	42.50	56.50	69.00
3	1	26.50	38.00	48.50	59.50
	2	30.00	40.00	50.00	61.00
	3	26.50	38.00	49.50	61.00
4	1	48.50	67.50	75.50	83.50
	2	46.50	62.50	73.50	83.50
	3	49.00	65.00	73.50	83.50
5	1	34.00	41.00	51.00	61.00
	2	34.50	44.50	55.50	67.00
	3	31.00	43.00	53.50	64.00
Total		506.50	696.00	868.00	1044.50

Analysis of the above data can be conducted as follows.

Step 1. Compute the total sum of squares using Equation (3.59) with values of Table 3.18.

$$SSTO = (29)^2 + (28)^2 + \ldots + (64)^2 - \frac{(3115.00)^2}{(4)(15)}$$

$$= 14961.58$$

Step 2. Construct an Isolate x PDA unit two-way table of totals by summing up the observations over different occasions and compute the marginal totals as shown in Table 3.19. Compute *SSG* and SSE_a using the values in this table and Equations (3.60) and (3.61).

Table 3.19. The Isolate x PDA unit totals computed from data in Table 3.18.

	Isolates					
PDA unit	1	2	3	4	5	Total
1	193.50	213.00	172.50	275.00	187.00	1041.00
2	190.50	205.50	181.00	266.00	201.50	1044.50
3	195.00	197.00	175.00	271.00	191.50	1029.50
Total	579.00	615.50	528.50	812.00	580.00	3115.00

$$SSG = \frac{(579.00)^2 + (615.50)^2 + \ldots + (580.00)}{(4)(3)} - \frac{(3115.00)^2}{(4)(15)}$$

$$= 4041.04$$

$$SSE_a = \frac{(193.50)^2 + (190.50)^2 + \ldots + (191.50)^2}{4} -$$

$$\frac{(579.00)^2 + (615.00)^2 + \ldots + (580.00)^2}{(4)(3)}$$

$$= 81.92$$

Step 3. Form the Isolate x Occasion two-way table of totals and compute the marginal totals as shown in Table 3.20. Compute *SSO*, *SSOG* and SSE_b using Equations (3.62) to (3.64).

Table 3.20. The Isolate x Occasion table of totals computed from data in Table 3.18

	Occasion				
Isolate	14 hr.	22 hr.	30 hr.	38 hr.	Total
1	86.00	123.00	164.00	206.00	579.00
2	94.00	133.50	173.50	214.50	615.50
3	83.00	116.00	148.00	181.50	528.50
4	144.00	195.00	222.50	250.50	812.00
5	99.50	128.50	160.00	192.00	580.00
Total	506.50	696.00	868.00	1044.50	3115.00

$$SSO = \frac{(506.50)^2 + (696.00)^2 + (868.00)^2 + (1044.50)^2}{15} - \frac{(3115.00)^2}{(4)(15)}$$

$$= 10637.08$$

$$SSOG = \frac{(86.00)^2 + (94.00)^2 + \ldots + (192.00)^2}{3}$$

$$- \frac{(579.00)^2 + (615.50)^2 + \ldots + (580.00)^2}{(4)(3)} - 10637.08$$

$$= 172.46$$

$$SSE_b = 14961.58 - 4041.04 - 81.92 - 10637.08 - 172.46$$

$$= 29.08$$

Step 4. Summarise the results as in Table 3.21 and perform the remaining calculations to obtain the mean squares and *F*-ratios using the equations reported in Table 3.17.

Table 3.21. ANOVA table for the data in Table 3.18.

Sources of variation	Degrees of freedom	Sum of squares	Mean sum of squares	*F*-ratio
Group	4	4041.04	1010.26	123.33*
Individuals within groups	10	81.92	8.19	
Occasion	3	10637.08	3545.69	3657.45*
Occasion x Group	12	172.46	14.37	14.82*
Occasion x Individuals within groups	30	29.08	0.97	
Total	59	14961.58		

Compare the computed values of *F* with tabular values of *F* with corresponding degrees of freedom at the desired probability level. All the computed *F* values in the above table are greater than corresponding tabular *F* values. Hence, we conclude that the variation due to groups, occasion and their interaction are significant meaning essentially that the isolates differ in their growth pattern across time.

4. DESIGN AND ANALYSIS OF EXPERIMENTS

Planning an experiment to obtain appropriate data and drawing inference out of the data with respect to any problem under investigation is known as *design and analysis of experiments.* This might range anywhere from the formulations of the objectives of the

experiment in clear terms to the final stage of the drafting reports incorporating the important findings of the enquiry. The structuring of the dependent and independent variables, the choice of their levels in the experiment, the type of experimental material to be used, the method of the manipulation of the variables on the experimental material, the method of recording and tabulation of data, the mode of analysis of the material, the method of drawing sound and valid inference etc. are all intermediary details that go with the design and analysis of an experiment.

4.1 Principles of experimentation

Almost all experiments involve the three basic principles, *viz*., randomization, replication and local control. These three principles are, in a way, complementary to each other in trying to increase the accuracy of the experiment and to provide a valid test of significance, retaining at the same time the distinctive features of their roles in any experiment. Before we actually go into the details of these three principles, it would be useful to understand certain generic terms in the theory experimental designs and also understand the nature of variation among observations in an experiment.

Before conducting an experiment, an *experimental unit* is to be defined. For example, a leaf, a tree or a collection of adjacent trees may be an experimental unit. An experimental unit is also sometimes referred as *plot*. A collection of plots is termed a *block*. Observations made on experimental units vary considerably. These variations are partly produced by the manipulation of certain variables of interest generally called *treatments*, built-in and manipulated deliberately in the experiment to study their influences. For instance, clones in clonal trials, levels and kinds of fertilizers in fertilizer trials etc. can be called treatments. Besides the variations produced in the observations due to these known sources, the variations are also produced by a large number of unknown sources such as uncontrolled variation in extraneous factors related to the environment, genetic variations in the experimental material other than that due to treatments, etc. They are there, unavoidable and inherent in the very process of experimentation. These variations because of their undesirable influences are called *experimental error* thereby meaning not an arithmetical error but variations produced by a set of unknown factors beyond the control of the experimenter.

It is further interesting to note that these errors introduced into the experimental observations by extraneous factors may be either *systematic* or *random* in their mode of incidence. The errors arising due to an equipment like a spring balance which goes out of calibration due to continued use or the error due to observer's fatigue are examples of systematic error. On the other hand, the unpredictable variation in the amount of leaves collected in litter traps under a particular treatment in a related experiment is random in nature. It is clear that any number of repeated measurements would not overcome systematic error where as it is very likely that the random errors would cancel out with repeated measurements. The three basic principle *viz*., randomization, replication and local control are devices to avoid the systematic error and to control the random error.

4.1.1. Randomization

Assigning the treatments or factors to be tested to the experimental units according to definite laws or probability is technically known as randomization. It is the randomization in its strict technical sense, that guarantees the elimination of systematic error. It further ensures that whatever error component that still persists in the observations is purely random in nature. This provides a basis for making a valid estimate of random fluctuations which is so essential in testing of significance of genuine differences.

Through randomization, every experimental unit will have the same chance of receiving any treatment. If, for instance, there are five clones of eucalyptus to be tried in say 25 plots, randomization ensures that certain clones will not be favoured or handicapped by extraneous sources of variation over which the experimenter has no control or over which he chooses not to exercise his control. The process of random allocation may be done in several ways, either by drawing lots or by drawing numbers from a page of random numbers, the page itself being selected at random. The method is illustrated in later sections dealing with individual forms of experimental designs.

4.1.2. Replication

Replication is the repetition of experiment under identical conditions but in the context of experimental designs, it refers to the number of distinct experimental units under the same treatment. Replication, with randomization, will provide a basis for estimating the error variance. In the absence of randomization, any amount of replication may not lead to a true estimate of error. The greater the number of replications, greater is the precision in the experiment.

The number of replications to be included in any experiment depends upon many factors like the homogeneity of experimental material, the number of treatments, the degree of precision required etc. As a rough rule, it may be stated that the number of replications in a design should provide at least 10 to 15 degrees of freedom for computing the experimental error variance.

4.1.3. Local control

Local control means the control of all factors except the ones about which we are investigating. Local control, like replication is yet another device to reduce or control the variation due to extraneous factors and increase the precision of the experiment. If, for instance, an experimental field is heterogeneous with respect of soil fertility, then the field can be divided into smaller blocks such that plots within each block tend to be more homogeneous. This kind of homogeneity of plots (experiment units) ensures an unbiased comparison of treatment means, as otherwise it would be difficult to attribute the mean difference between two treatments solely to differences between treatments when the plot differences also persist. This type of local control to achieve

homogeneity of experimental units, will not only increase the accuracy of the experiment, but also help in arriving at valid conclusions.

In short, it may be mentioned that while randomization is a method of eliminating a systematic error (*i.e.*, bias) in allocation thereby leaving only random error component of variation, the other two *viz*., replication and local control try to keep this random error as low as possible. All the three however are essential for making a valid estimate of error variance and to provide a valid test of significance.

4.2. Completely randomized design

A completely randomized design (CRD) is one where the treatments are assigned completely at random so that each experimental unit has the same chance of receiving any one treatment. For the CRD, any difference among experimental units receiving the same treatment is considered as experimental error. Hence, CRD is appropriate only for experiments with homogeneous experimental units, such as laboratory experiments, where environmental effects are relatively easy to control. For field experiments, where there is generally large variation among experimental plots in such environmental factors as soil, the CRD is rarely used.

4.2.1. Layout

The step-by-step procedure for randomization and layout of a CRD are given here for a pot culture experiment with four treatments A, B, C and D, each replicated five times.

Step 1. Determine the total number of experimental plots (n) as the product of the number of treatments (t) and the number of replications (r); that is, $n = rt$. For our example, $n = 5 \times 4 = 20$. Here, one pot with a single plant in it may be called a plot. In case the number of replications is not the same for all the treatments, the total number of experimental pots is to be obtained as the sum of the replications for each treatment. *i.e.*,

$$n = \sum_{i=1}^{t} r_i$$ where r_i is the number of times the ith treatment replicated

Step 2. Assign a plot number to each experimental plot in any convenient manner; for example, consecutively from 1 to n.

Step 3. Assign the treatments to the experimental plots randomly using a table of random numbers as follows. Locate a starting point in a table of random numbers (Appendix 6) by closing your eyes and pointing a finger to any position in a page. For our example, the starting point is taken at the intersection of the sixth row and the twelfth (single) column of two-digit numbers. Using the starting point obtained, read downward vertically to obtain $n = 20$ distinct two-digit random numbers. For our example, starting at the intersection of the sixth

row and the twelfth column, the 20 distinct two-digit random numbers are as shown here together with their corresponding sequence of appearance.

Random number: 37, 80, 76, 02, 65, 27, 54, 77, 48, 73,
Sequence : 1, 2, 3, 4, 5, 6, 7, 8, 9, 10,

Random number: 86, 30, 67, 05, 50, 31, 04, 18, 41, 89
Sequence : 11, 12, 13, 14, 15, 16, 17, 18, 19, 20

Rank the *n* random numbers obtained in ascending or descending order. For our example, the 20 random numbers are ranked from the smallest to the largest, as shown in the following:

Random Number	Sequence	Rank	Random Number	Sequence	Rank
37	1	8	86	11	19
80	2	18	30	12	6
76	3	16	67	13	14
02	4	1	05	14	3
65	5	13	50	15	11
27	6	5	31	16	7
54	7	12	04	17	2
77	8	17	18	18	4
48	9	10	41	19	9
73	10	15	89	20	20

Divide the *n* ranks derived into *t* groups, each consisting of *r* numbers, according to the sequence in which the random numbers appeared. For our example, the 20 ranks are divided into four groups, each consisting of five numbers, as follows:

Group Number	Ranks in the Group				
1	8	13	10	14	2
2	18	5	15	3	4
3	16	12	19	11	9
4	1	17	6	7	20

Assign the *t* treatments to the *n* experimental plots, by using the group number as the treatment number and the corresponding ranks in each group as the plot number in which the corresponding treatment is to be assigned. For our example, the first group is assigned to treatment A and plots numbered 8, 13, 10, 14 and 2 are assigned to receive this treatment; the second group is assigned to treatment B with plots numbered 18, 5, 15, 3 and 4; the third group is assigned to treatment C with plots numbered 16, 12, 19, 11 and 9; and the fourth group to treatment D with plots numbered 1, 17, 6, 7 and 20. The final layout of the experiment is shown Figure 4.1.

Plot no → / Treatment →			
1 D	2 A	3 B	4 B
5 B	6 D	7 D	8 A
9 C	10 A	11 C	12 C
13 A	14 A	15 B	16 C
17 D	18 B	19 C	20 D

Figure 4.1. A sample layout of a completely randomised design with four treatments (A, B, C and D) each replicated five times

4.2.2. Analysis of variance

There are two sources of variation among the *n* observations obtained from a CRD trial. One is the variation due to treatments, the other is experimental error. The relative size of the two is used to indicate whether the observed difference among treatments is real or is due to chance. The treatment difference is said to be real if treatment variation is sufficiently larger than experimental error.

A major advantage of the CRD is the simplicity in the computation of its analysis of variance, especially when the number of replications is not uniform for all treatments. For most other designs, the analysis of variance becomes complicated when the loss of data in some plots results in unequal replications among the treatments tested.

The steps involved in the analysis of variance for data from a CRD experiment with unequal number of replications are given below. The formulae are easily adaptable to the case of equal replications and hence not shown separately. For illustration, data from a laboratory experiment are used, in which observations were made on mycelial growth of different *Rizoctonia solani* isolates on PDA medium (Table 4.1).

Step 1. Group the data by treatments and calculate the treatment totals (T_i) and grand total (G). For our example, the results are shown in Table 4.1 itself.

Step 2. Construct an outline of ANOVA table as in Table 4.2.

Table 4.1. Mycelial growth in terms of diameter of the colony (mm) of *R. solani* isolates on PDA medium after 14 hours of incubation.

R. solani	Mycelial growth	Treatment	Treatment

isolates				total	mean
	Repl. 1	Repl. 2	Repl. 3	(T_i)	
RS 1	29.0	28.0	29.0	86.0	28.67
RS 2	33.5	31.5	29.0	94.0	31.33
RS 3	26.5	30.0		56.5	28.25
RS 4	48.5	46.5	49.0	144.0	48.00
RS 5	34.5	31.0		65.5	32.72
Grand total				446.0	
Grand mean					34.31

Table 4.2. Schematic representation of ANOVA of CRD with unequal replications.

Source of variation	Degree of freedom (*df*)	Sum of squares (*SS*)	Mean square $\left(MS = \frac{SS}{df}\right)$	Computed *F*
Treatment	t - 1	*SST*	*MST*	$\frac{MST}{MSE}$
Error	n - t	*SSE*	*MSE*	
Total	n - 1	*SSTO*		

Step 3. With the treatment totals (T_i) and the grand total (G) of Table 4.1, compute the correction factor and the various sums of squares, as follows. Let y_{ij} represent the observation on the jth PDA medium belonging to the ith isolate; i = 1, 2, …, t ; j = 1, 2, …, $r_{i.}$.

$$C.\ F. = \frac{G^2}{n} \tag{4.1}$$

$$= \frac{(446)^2}{13}$$

$$= 15301.23$$

$$SSTO = \sum_{i=1}^{t}\sum_{j=1}^{r_i} y_{ij}^2 - C.\ F. \tag{4.2}$$

$$= \left[(29.0)^2 + (28.0)^2 + \ldots + (31.0)^2\right] - 15301.23$$

$$= 789.27$$

$$SST = \sum_{i=1}^{t} \frac{T_i^2}{r_i} - C.\ F. \tag{4.3}$$

$$= \left[\frac{(86)^2}{3} + \frac{(94)^2}{3} + ... + \frac{(65.5)^2}{2}\right] - 15301.23$$

$$= 762.69$$

$$SSE = SSTO - SST \qquad (4.4)$$

$$= 789.27 - 762.69 = 26.58$$

Step 4. Enter all the values of sums of squares in the ANOVA table and compute the mean squares and F value as shown in the Table 4.2.

Step 5. Obtain the tabular F values from Appendix 3, with f_1 and f_2 degrees of freedom where f_1 = treatment df = (t - 1) and f_2 = error df = (n - t), respectively. For our example, the tabular F value with f_1 = 4 and f_2 = 8 degrees of freedom is 3.84 at 5% level of significance. The above results are shown in Table 4.3.

Table 4.3. ANOVA of mycelial growth data of Table 4.1.

Source of variation	Degree of freedom	Sum of squares	Mean square	Computed F	Tabular F 5%
Treatment	4	762.69	190.67	57.38*	3.84
Error	8	26.58	3.32		
Total	12	789.27			

* Significant at 5% level

Step 7. Compare the computed F value of Step 4 with the tabular F value of Step 5, and decide on the significance of the difference among treatments using the following rules:

(i) If the computed F value is larger than the tabular F value at 5% level of significance, the variation due to treatments is said to be *significant*. Such a result is generally indicated by placing an asterisk on the computed F value in the analysis of variance.

(ii) If the computed F value is smaller than or equal to the tabular F value at 5% level of significance, the variation due to treatments is said to be *nonsignificant*. Such a result is indicated by placing *ns* on the computed F value in the analysis of variance or by leaving the F value without any such marking.

Note that a nonsignificant F in the analysis of variance indicates the failure of the experiment to detect any differences among treatments. It does not, in any way, prove that all treatments are the same, because the failure to detect treatment differences based on the nonsignificant F test, could be the result of either a very small or no difference among the treatments or due to large experimental error, or both. Thus, whenever the F test is nonsignificant, the researcher should examine the size of the experimental error and the numerical

differences among the treatment means. If both values are large, the trial may be repeated and efforts made to reduce the experimental error so that the differences among treatments, if any, can be detected. On the other hand, if both values are small, the differences among treatments are probably too small to be of any economic value and, thus, no additional trials are needed.

For our example, the computed *F* value of 57.38 is larger than the tabular *F* value of 3.84 at the 5% level of significance. Hence, the treatment differences are said to be significant. In other words, chances are less than 5 in 100 that all the observed differences among the five treatment means could be due to chance. It should be noted that such a significant *F* test verifies the existence of some differences among the treatments tested but does not specify the particular pair (or pairs) of treatments that differ significantly. To obtain this information, procedures for comparing treatment means, discussed in Section 4.2.3. are needed.

Step 8. Compute the grand mean and the coefficient of variation (*cv*) as follows:

$$\text{Grand mean} = \frac{G}{n} \tag{4.5}$$

$$cv = \frac{\sqrt{MSE}}{\text{Grand mean}}(100) \tag{4.6}$$

For our example,

$$\text{Grand mean} = \frac{446}{13} = 34.31$$

$$cv = \frac{\sqrt{3.32}}{34.31}(100) = 5.31\%$$

The *cv* affects the degree of precision with which the treatments are compared and is a good index of the reliability of the experiment. It is an expression of the overall experimental error as percentage of the overall mean; thus, the higher the *cv* value, the lower is the reliability of the experiment. The *cv* varies greatly with the type of experiment, the crop grown, and the characters measured. An experienced researcher, however, can make a reasonably good judgement on the acceptability of a particular *cv* value for a given type of experiment. Experimental results having a *cv* value of more than 30 % are to be viewed with caution.

4.2.3. Comparison of treatments

One of the most commonly used test procedures for pair comparisons in forestry research is the least significant difference (LSD) test. Other test procedures, such as

Duncan's multiple range test (DMRT), the honestly significant difference (HSD) test and the Student-Newman-Keuls range test, can be found in Gomez and Gomez (1980), Steel and Torrie (1980) and Snedecor and Cochran (1980). The LSD test is described in the following.

The LSD test is the simplest of the procedures for making pair comparisons. The procedure provides for a single LSD value, at a prescribed level of significance, which serves as the boundary between significant and nonsignificant difference between any pair of treatment means. That is, two treatments are declared significantly different at a prescribed level of significance if their difference exceeds the computed LSD value, otherwise they are not considered significantly different.

The LSD test is most appropriate for making planned pair comparisons but, strictly speaking, is not valid for comparing all possible pairs of means, especially when the number of treatments is large. This is so because the number of possible pairs of treatment means increases rapidly as the number of treatments increases. The probability that, due to chance alone, at least one pair will have a difference that exceeds the LSD value increases with the number of treatments being tested. For example, in experiments where no real difference exists among all treatments, it can be shown that the numerical difference between the largest and the smallest treatment means is expected to exceed the LSD value at the 5% level of significance 29% of the time when 5 treatments are involved, 63% of the time when 10 treatments are involved and 83% of the time when 15 treatments are involved. Thus one must avoid use of the LSD test for comparisons of all possible pairs of means. If the LSD test must be used, apply it only when the *F* test for treatment effect is significant and the number of treatments is not too large, say, less than six.

The procedure for applying the LSD test to compare any two treatments, say the *i*th and the *j*th treatments, involves the following steps:

Step 1. Compute the mean difference between the *i*th and the *j*th treatment as:

$$d_{ij} = \bar{y}_i - \bar{y}_j \qquad (4.7)$$

where $\bar{y}_i$ and $\bar{y}_j$ are the means of the *i*th and the *j*th treatments.

Step 2. Compute the LSD value at α level of significance as:

$$\text{LSD}_\alpha = \left(t_{v;\,\alpha}\right)\left(s_{\bar{d}}\right) \qquad (4.8)$$

where $s_{\bar{d}}$ is the standard error of the mean difference and $t_{v;\,\alpha}$ is the Student's *t* value, from Appendix 2, at α level of significance and with v = Degrees of freedom for error.

Step 3. Compare the mean difference computed in Step 1 with the LSD value computed in Step 2 and declare the *i*th and *j*th treatments to be significantly different at the α level of significance, if the absolute value of d_{ij} is greater than the LSD value.

In applying the foregoing procedure, it is important that the appropriate standard error of the mean difference ($s_{\bar{d}}$)for the treatment pair being compared is identified. This task is affected by the experimental design used, the number of replications of the two treatments being compared, and the specific type of means to be compared. In the case of CRD, when the two treatments do not have the same number of replications, $s_{\bar{d}}$ is computed as:

$$s_{\bar{d}} = \sqrt{s^2\left(\frac{1}{r_i} + \frac{1}{r_j}\right)} \qquad (4.9)$$

where r_i and r_j are the number of replications of the *i*th and the *j*th treatments and s^2 is the error mean square in the analysis of variance.

As an example, use the data from Table 4.1. The researcher wants to compare the five isolates of *R. solani*, with respect to the mycelial growth on PDA medium. The steps involved in applying the LSD test would be the following.

Step 1. Compute the mean difference between each pair of treatments (isolates) as shown in Table 4.4.

Step 2. Compute the LSD value at α level of significance. Because some treatments have three replications and others have two, three sets of LSD values must be computed.

For comparing two treatments each having three replications, compute the LSD value as follows.

$$\text{LSD}_{.05} = 2.31\sqrt{\frac{2(3.32)}{3}} = 3.44 \text{ mm}$$

where the value of $s^2 = 3.32$ is obtained from Table 4.3 and the Student's *t* value of 2.31 for 8 degrees of freedom at 5% level is obtained from Appendix 2.

For comparing two treatments each having three replications, compute the LSD value as follows.

$$\text{LSD}_{.05} = 2.31\sqrt{\frac{2(3.32)}{2}} = 4.21 \text{ mm}$$

For comparing two treatments one having two replications and the other having three replications, the LSD value is,

$$\text{LSD}_{.05} = 2.31\sqrt{3.32(1/3 + 1/2)}$$
$$= 3.84 \text{ mm}$$

Step 3. Compare difference between each pair of treatments computed in Step 1 to the corresponding LSD values computed in Step 2 and place the appropriate

asterisk notation. For example, the mean difference between the first treatment (with three replications) and the second treatment (with three replications) is 2.66 mm. Since the mean difference is less than the corresponding LSD value of 3.44 mm it is declared to be nonsignificant at 5% level of significance. On the other hand, the mean difference between the first treatment (with three replications) and the second treatment (with two replications) is, 4.05 mm. Since the mean difference is higher than the corresponding LSD value of 3.84, it is declared to be significant at the 5% level and is indicated with asterisks. The test results for all pairs of treatments are given in Table 4.4.

Table 4.4. Comparison between mean diameter (mm) of each pair of treatments using the LSD test with unequal replications, for data in Table 4.1.

Treatment	RS 1	RS 2	RS 3	RS 4	RS 5
RS 1	0.00	2.66 (3.44)	0.42 (3.84)	19.33* (3.44)	4.05* (3.84)
RS 2		0.00	3.08 (3.84)	16.67* (3.44)	1.39 (3.84)
RS 3			0.00	19.75* (3.84)	4.47* (4.21)
RS 4				0.00	15.28* (3.84)
RS 5					0.00

* Significant at 5% level

Note: The values in the parenthesis are LSD values

Before leaving this section, one point useful in deciding the number of replications required in an experiment for achieving reasonable level of reliability is mentioned here. As indicated earlier, one thumb rule is to take that many replications which will make the error degrees of freedom around 12. The idea behind this rule is that critical values derived from some of the distributions like Student's t or F almost stabilize after 12 degrees of freedom thereby providing some extent of stability to the conclusions drawn from such experiments. For instance, if one were to plan a CRD with equal replications for t treatments, one would equate the error df of $t(r\text{-}1)$ to 12 and solve for r for known values of t. Similar strategies can be followed for many other designs also that are explained in later sections.

4.3. Randomized complete block design

The randomized complete block design (RCBD) is one of the most widely used experimental designs in forestry research. The design is especially suited for field experiments where the number of treatments is not large and there exists a conspicuous factor based on which homogenous sets of experimental units can be identified. The

primary distinguishing feature of the RCBD is the presence of blocks of equal size, each of which contains all the treatments.

4.3.1 Blocking technique

The purpose of blocking is to reduce the experimental error by eliminating the contribution of known sources of variation among the experimental units. This is done by grouping the experimental units into blocks such that variability within each block is minimized and variability among blocks is maximized. Since only the variation within a block becomes part of the experimental error, blocking is most effective when the experimental area has a predictable pattern of variability.

An ideal source of variation to use as the basis for blocking is one that is large and highly predictable. An example is soil heterogeneity, in a fertilizer or provenance trial where yield data is the primary character of interest. In the case of such experiments, after identifying the specific source of variability to be used as the basis for blocking, the size and the shape of blocks must be selected to maximize variability among blocks. The guidelines for this decision are (i) When the gradient is unidirectional (*i.e.*, there is only one gradient), use long and narrow blocks. Furthermore, orient these blocks so that their length is perpendicular to the direction of the gradient. (ii) When the fertility gradient occurs in two directions with one gradient much stronger than the other, ignore the weaker gradient and follow the preceding guideline for the case of the unidirectional gradient. (iii) When the fertility gradient occurs in two directions with both gradients equally strong and perpendicular to each other, use blocks that are as square as possible or choose other designs like *latin square design* (Gomez and Gomez, 1980).

Whenever blocking is used, the identity of the blocks and the purpose for their use must be consistent throughout the experiment. That is, whenever a source of variation exists that is beyond the control of the researcher, it should be ensured that such variation occurs among blocks rather than within blocks. For example, if certain operations such as application of insecticides or data collection cannot be completed for the whole experiment in one day, the task should be completed for all plots of the same block on the same day. In this way, variation among days (which may be enhanced by weather factors) becomes a part of block variation and is, thus, excluded from the experimental error. If more than one observer is to make measurements in the trial, the same observer should be assigned to make measurements for all plots of the same block. This way, the variation among observers if any, would constitute a part of block variation instead of the experimental error.

4.3.2. Layout

The randomization process for a RCBD is applied separately and independently to each of the blocks. The procedure is illustrated for the case of a field experiment with six treatments A, B, C, D, E, F and three replications.

Step1. Divide the experimental area into *r* equal blocks, where *r* is the number of replications, following the blocking technique described in Section 4.3.1. For our example, the experimental area is divided into three blocks as shown in Figure 4.2. Assuming that there is a unidirectional fertility gradient along the length of the experimental field, block shape is made rectangular and perpendicular to the direction of the gradient.

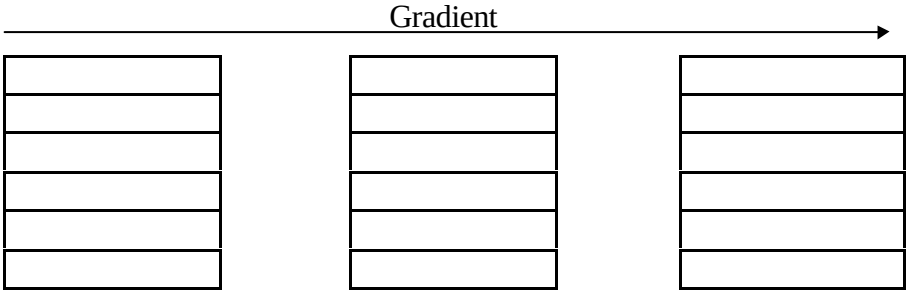


Figure 4.2. Division of an experimental area into three blocks, each consisting of six plots, for a randomized complete block design with six treatments and three replications. Blocking is done such that blocks are rectangular and perpendicular to the direction of the unidirectional gradient (indicated by the arrow).

Step2. Subdivide the first block into *t* experimental plots, where *t* is the number of treatments. Number the *t* plots consecutively from 1 to *t*, and assign *t* treatments at random to the *t* plots following any of the randomization schemes for the CRD described in Section 4.2.1. For our example, block I is subdivided into six equisized plots, which are numbered consecutively from top to bottom. (Figure 4.3) and the six treatments are assigned at random to the six plots using the table of random numbers as follows:

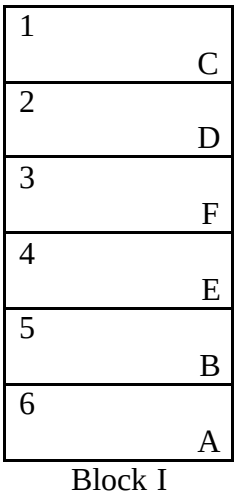


Figure 4.3. Plot numbering and random assignment of six treatments (A, B, C, D, E, and F) to the six plots of Block I.

Step 3. Repeat Step 2 completely for each of the remaining blocks. For our example, the final layout is shown in Figure 4.4.

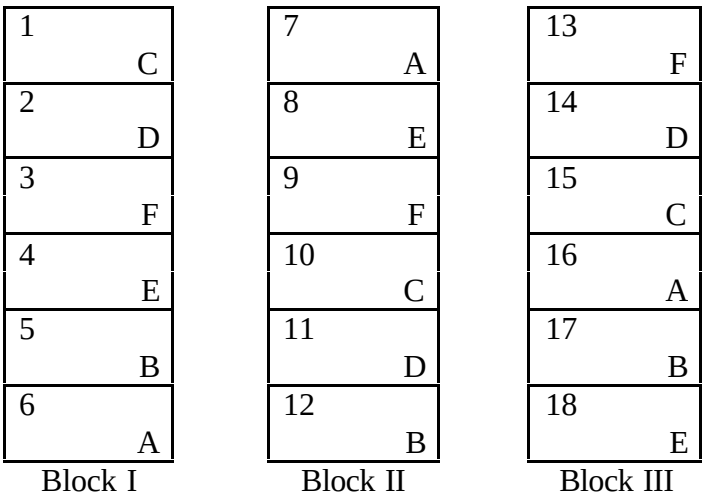


Figure 4.4. A sample layout of a randomized complete block design with six treatments (A, B, C, D, E and F) and three replications.

4.3.3. Analysis of variance

There are three sources of variability in a RCBD : treatment, replication (or block) and experimental error. Note that this is one more than that for a CRD, because of the addition of replication, which corresponds to the variability among blocks.

To illustrate the steps involved in the analysis of variance for data from a RCBD, data from an experiment is made use of, wherein eight provenances of *Gmelina arborea* were compared with respect to the girth at breast-height (gbh) of the trees attained since 6 years of planting (Table 4.5).

Table 4.5. Mean gbh (cm) of trees in plots of provenances of *Gmelina arborea*, 6 years after planting, in a field experiment laid out under RCBD.

Treatment (Provenance)	Replication			Treatment total (T_i)	Treatment mean
	I	II	III		
1	30.85	38.01	35.10	103.96	34.65
2	30.24	28.43	35.93	94.60	31.53
3	30.94	31.64	34.95	97.53	32.51
4	29.89	29.12	36.75	95.76	31.92
5	21.52	24.07	20.76	66.35	22.12
6	25.38	32.14	32.19	89.71	29.90
7	22.89	19.66	26.92	69.47	23.16
8	29.44	24.95	37.99	92.38	30.79
Rep. total (R_j)	221.15	228.02	260.59		
Grand total (G)				709.76	
Grand mean					29.57

Step 1. Group the data by treatments and replications and calculate treatment totals (T_i), replication totals (R_j) and grand total (G), as shown in Table 4.5.

Step 2. Construct the outline of the analysis of variance as follows:

Table 4.6. Schematic representation of ANOVA of RCBD

Source of variation	Degree of freedom (df)	Sum of squares (SS)	Mean square $\left(MS = \frac{SS}{df}\right)$	Computed F
Replication	r - 1	SSR	MSR	
Treatment	t - 1	SST	MST	$\frac{MST}{MSE}$
Error	(r - 1)(t - 1)	SSE	MSE	
Total	rt - 1	$SSTO$		

Step 3. Compute the correction factor and the various sums of squares (SS) given in the above table as follows. Let y_{ij} represent the observation made from jth block on the ith treatment; $i = 1,\ldots,t$; $j = 1,\ldots,r$.

$$CF = \frac{G^2}{rt} \tag{4.10}$$

$$= \frac{(709.76)^2}{(3)(8)} = 20989.97$$

$$SSTO = \sum_{i=1}^{t}\sum_{j=1}^{r} y_{ij}^2 - C.F. \tag{4.11}$$

$$= \left[(30.85)^2 + (38.01)^2 + \ \ldots\ + (37.99)^2\right] - \ 20989.97$$

$$= \ 678.42$$

$$SSR = \frac{\sum_{j=1}^{r} R_j^2}{t} - C.F. \tag{4.12}$$

$$= \frac{(221.15)^2 + (228.02)^2 + (260.59)^2}{8} - \ 20989.97$$

$$= 110.98$$

$$SST = \frac{\sum_{i=1}^{t} T_i^2}{r} - C.F. \tag{4.13}$$

$$= \frac{(103.96)^2 + (94.60)^2 + \ \ldots\ + (92.38)^2}{3} - 20989.97$$

$$= 426.45$$

$$SSE = SSTO - SSR - SST \quad (4.14)$$
$$= 678.42 - 110.98 - 426.45 = 140.98$$

Step 4. Using the values of sums of squares obtained, compute the mean square and the *F* value for testing the treatment differences as shown in the Table 4.6. The results are shown in Table 4.7.

Table 4.7 ANOVA of gbh data in Table 4.5.

Source of variation	Degree of freedom	Sum of Squares	Mean Square	Computed *F*	Tabular *F* 5%
Replication	2	110.98	55.49		
Treatment	7	426.45	60.92	6.05*	2.76
Error	14	140.98	10.07		
Total	23	678.42			

*Significant at 5% level

Step 5. Obtain the tabular *F* values from Appendix 3, for f_1 = treatment *df* and f_2 = error *df*. For our example, the tabular *F* value for f_1 = 7 and f_2 = 14 degrees of freedom is 2.76 at the 5% level of significance.

Step 6. Compare the computed *F* value of step 4 with the tabular *F* values of step 5, and decide on the significance of the differences among treatments. Because the computed *F* value of 6.05 is greater than the tabular *F* value at the 5% level of significance, we conclude that the experiment shows evidence the existence of significant differences among the provenances with respect to their growth in terms gbh.

Step 7. Compute the coefficient of variation as:

$$cv = \frac{\sqrt{\text{Error } MS}}{\text{GrandMean}}(100) \quad (4.15)$$
$$= \frac{\sqrt{10.37}}{29.57}(100) = 10.89\%$$

The relatively low value of *cv* indicates the reasonable level of precision attained in the field experiment.

4.3.4. Comparison of treatments

The treatment means are compared as illustrated for the case of CRD in Section 4.2.3. using the formulae,

$$LSD_{\alpha} = \left(t_{v;\,\alpha}\right)\left(s_{\bar{d}}\right) \quad (4.16)$$

where $s_{\bar{d}}$ is the standard error of the difference between treatment means and $t_{v;\,a}$ is the tabular *t* value, from Appendix 2, at α level of significance and with v = Degrees of freedom for error. The quantity $s_{\bar{d}}$ is computed as:

$$s_{\bar{d}} = \sqrt{\frac{2s^2}{r}} \qquad (4.17)$$

where s^2 is the mean square due to error and *r* is the number of replications.

For illustration, the analysis carried out on data given in Table 4.5 is continued to compare all the possible pairs of treatments through LSD test.

Step 1. Compute the difference between treatment means as shown in Table 4.8.

Table 4.8. Difference between mean gbh (cm) for each pair of treatments of data in Table 4.4.

Treatment	1	2	3	4	5	6	7	8
1	0.00	3.12	2.14	2.73	12.53*	4.75	11.49*	3.86
2		0.00	0.98	0.39	9.41*	1.63	8.37*	0.74
3			0.00	0.59	10.39*	2.61	9.35*	1.72
4				0.00	9.8*	2.02	8.76*	1.13
5					0.00	7.78*	1.04	8.67*
6						0.00	6.74*	0.89
7							0.00	7.63*
8								0.00

* Significant at 5% level

Step 2.Compute the LSD value at α level of significance. Since all the treatments are equally replicated, we need to compute only one LSD value. The LSD value is computed using Equations (4.16) and (4.17).

$$\text{LSD}_{.05} = 2.14\sqrt{\frac{2(10.07)}{3}} = 5.54 \text{ cm}$$

Step 3.Compare difference among the treatment means against the computed value of LSD and place the asterisk against significant differences. The results are shown in Table 4.8.

4.3.5. Estimation of missing values

A missing data situation occurs whenever a valid observation is not available for any one of the experimental units. Missing data could occur due to accidental improper application of treatments, erroneous observations, destruction of experimental units due

to natural calamities like fire, damage due to wildlife etc. It is extremely important, however, to carefully examine the reasons for missing data. The destruction of the experimental material must not be the result of the treatment effect. If a plot has no surviving plants because it has been grazed by stray cattle or vandalized by thieves, each of which is clearly not treatment related, missing data should be appropriately declared. On the other hand, for example, if a control plot *(i.e.*, untreated plot) in an insecticide trial is totally damaged by the insects, the destruction is a logical consequence of that plot being the control plot. Thus, the corresponding plot data should be accepted as valid (*i.e.*, zero yield if all plants in the plot are destroyed, or the actual low yield value if some plants survive) instead of treating it as missing data.

Occurrence of missing data results in two major difficulties; loss of information and non- applicability of the standard analysis of variance. When an experiment has one or more observations missing, the standard computational procedures of the analysis of variance no longer apply except for CRD. One alternative in such cases is the use of the *missing data formula technique.* In the missing data formula technique, an estimate of a single missing observation is provided through an appropriate formula according to the experimental design used. This estimate is used to replace the missing data and the augmented data set is then subjected, with some slight modifications, to the standard analysis of variance.

It is to be noted that an estimate of the missing data obtained through the missing data formula technique does not supply any additional information, the data once lost is not retrievable through any amount of statistical manipulation. What the procedure attempts to do is to allow the researcher to compute the analysis of variance in the usual manner (*i.e.*, as if the data were complete) without resorting to the more complex procedures needed for incomplete data sets.

A single missing value in a randomized complete block design is estimated as:

$$y = \frac{rB_0 + tT_0 - G_0}{(r-1)(t-1)} \qquad (4.18)$$

where y = Estimate of missing data
t = Number of treatments
r = Number of replications
B_0 = Total of observed values of the replication that contains the missing data
T_0 = Total of observed values of the treatment that contains the missing data
G_0 = Grand total of all observed values

The missing data is replaced by the computed value of y and the usual computational procedure of the analysis of variance is applied to the augmented dataset with some modifications.

The procedure is illustrated with data of Table 4.5, with the value of the sixth treatment (sixth provenance) in replication II assumed to be missing, as shown in Table 4.9. The

steps in the computation of the analysis of variance and pair comparisons of treatment means are as follows.

Step 1. Firstly, estimate the missing value, using Equation (4.18) and the values of totals in Table 4.9.

$$y = \frac{3(195.88) + 8(57.57) - 677.62}{(3-1)(8-1)} = 26.47$$

Table 4.9. Data of Table 4.5 with one missing observation

Treatment (Provenance)	Replication			Treatment total (T)
	Rep. I	Rep II	Rep. III	
1	30.85	38.01	35.1	103.96
2	30.24	28.43	35.93	94.6
3	30.94	31.64	34.95	97.53
4	29.89	29.12	36.75	95.76
5	21.52	24.07	20.76	66.35
6	25.38	M	32.19	(57.57=T_0)
7	22.89	19.66	26.92	69.47
8	29.44	24.95	37.99	92.38
Rep. total (R)	221.15	(195.88=B_0)	260.59	
Grand total (G)				(677.62=G_0)

M = Missing data

Step 2. Replace the missing data of Table 4.9. by its estimated value computed in step 1, as shown in Table 4.10 and carry out the analysis of variance of the augmented data set based on the standard procedure of Section 4.3.3.

Table 4.10. Data in Table 4.7 with the missing data replaced by the value estimated from the missing data formula technique.

Treatment (Provenance)	Replication			Treatment total (T)
	Rep. I	Rep II	Rep. III	
1	30.85	38.01	35.1	103.96
2	30.24	28.43	35.93	94.6
3	30.94	31.64	34.95	97.53
4	29.89	29.12	36.75	95.76
5	21.52	24.07	20.76	66.35
6	25.38	26.47[a]	32.19	84.04
7	22.89	19.66	26.92	69.47
8	29.44	24.95	37.99	92.38
Rep. total (R)	221.15	222.35	260.59	
Grand total (G)				704.09

[a] Estimate of the missing data obtained from missing data formula technique

Step 3. Make the following modifications to the analysis of variance obtained in Step 2; Subtract 1 from both the total and error *df*. For our example, the total *df* of 23 becomes 22 and the error *df* of 14 becomes 13. Compute the correction factor for bias (*B*) as,

$$B = \frac{\left[B_0 - (t-1)y\right]^2}{t(t-1)} \qquad (4.19)$$

$$= \frac{\left[195.88 - (8-1)(26.47)\right]^2}{8(8-1)}$$

$$= 2.00$$

and subtract the computed *B* value of 2.00 from the treatment sum of squares and the total sum of squares. For our example, the *SSTO* and the *SST*, computed in Step 2 from the augmented data of Table 4.10, are 680.12 and 432.09, respectively. Subtracting the *B* value of 2.00 from these *SS* values, we obtain the adjusted *SST* and the adjusted *SSTO* as:

Adjusted *SST* = 432.09 - 2.00
= 430.09

Adjusted *SSTO* = 680.12 - 2.00
= 678.12

The resulting ANOVA is shown in Table 4.11.

Table 4.11. Analysis of variance of data in Table 4.7 with one missing value estimated by the missing data formula technique.

Source of variation	Degree of freedom	Sum of squares	Mean square	Computed *F*	Tabular *F* 5 %
Replication	2	125.80	62.90	6.69	
Treatment	7	430.09	61.44	6.53*	2.83
Error	13	122.23	9.40		
Total	22	678.12			

* Significant at 5% level of significance

Step 4. For pairwise comparisons of treatment means where one of the treatments has missing data, compute the standard error of the mean difference $s_{\bar{d}}$ as:

$$s_{\bar{d}} = \sqrt{s^2\left[\frac{2}{r} + \frac{t}{r(r-1)(t-1)}\right]} \qquad (4.20)$$

where s^2 is the error mean square from the analysis of variance of Step 3, *r* is the number of replications, and *t* is the number of treatments.

For example, to compare the mean of the sixth treatment (the treatment with missing data) with any one of the other treatments, $s_{\bar{d}}$ is computed as:

$$s_{\bar{d}} = \sqrt{9.40\left[\frac{2}{3} + \frac{8}{(3)(2)(7)}\right]} = 2.84$$

This computed $s_{\bar{d}}$ is appropriate for use in the computation of the LSD values. For illustration, the computation of the LSD values is shown below. Using t_v as the tabular *t* value for 13 *df* at 5% level of significance, obtained from Appendix 3, the LSD values for comparing the sixth treatment mean with any other treatment mean is computed as:

$$LSD_{\alpha} = t_{v;\,a}\ s_{\bar{d}} \qquad (4.21)$$

$$LSD_{.05} = (2.16)(2.84) = 6.13$$

4.4. Factorial experiments

Response variable(s) in any experiment can be found to be affected by a number of factors in the overall system some of which are controlled or maintained at desired levels in the experiment. An experiment in which the treatments consist of all possible combinations of the selected levels in two or more factors is referred as a factorial experiment. For example, an experiment on rooting of cuttings involving two factors, each at two levels, such as two hormones at two doses, is referred to as a 2 x 2 or a 2^2 factorial experiment. Its treatments consist of the following four possible combinations of the two levels in each of the two factors.

Treatment number	Treatment combination	
	Hormone	Dose (ppm)
1	NAA	10
2	NAA	20
3	IBA	10
4	IBA	20

The term *complete factorial experiment* is sometimes used when the treatments include all combinations of the selected levels of the factors. In contrast, the term fractional factorial experiment is used when only a fraction of all the combinations is tested. Throughout this manual, however, complete factorial experiments are referred simply as factorial experiments. Note that the term *factorial* describes a specific way in which the treatments are formed and does not, in any way, refer to the design used for laying out the experiment. For example, if the foregoing 2^2 factorial experiment is in a

randomized complete block design, then the correct description of the experiment would be 2^2 *factorial experiment in randomized complete block design.*

The total number of treatments in a factorial experiment is the product of the number of levels of each factor; in the 2^2 factorial example, the number of treatments is 2 x 2 = 4, in the 2^3 factorial, the number of treatments is 2 x 2 x 2 = 8. The number of treatments increases rapidly with an increase in the number of factors or an increase in the levels in each factor. For a factorial experiment involving 5 clones, 4 espacements, and 3 weed-control methods, the total number of treatments would be 5 x 4 x 3 = 60. Thus, indiscriminate use of factorial experiments has to be avoided because of their large size, complexity, and cost. Furthermore, it is not wise to commit oneself to a large experiment at the beginning of the investigation when several small preliminary experiments may offer promising results. For example, a tree breeder has collected 30 new clones from a neighbouring country and wants to assess their reaction to the local environment. Because the environment is expected to vary in terms of soil fertility, moisture levels, and so on, the ideal experiment would be one that tests the 30 clones in a factorial experiment involving such other variable factors as fertilizer, moisture level, and population density. Such an experiment, however, becomes extremely large as factors other than clones are added. Even if only one factor, say nitrogen or fertilizer with three levels were included, the number of treatments would increase from 30 to 90. Such a large experiment would mean difficulties in financing, in obtaining an adequate experimental area, in controlling soil heterogeneity, and so on. Thus, the more practical approach would be to test the 30 clones first in a single-factor experiment, and then use the results to select a few clones for further studies in more detail. For example, the initial single-factor experiment may show that only five clones are outstanding enough to warrant further testing. These five clones could then be put into a factorial experiment with three levels of nitrogen, resulting in an experiment with 15 treatments rather than the 90 treatments needed with a factorial experiment with 30 clones.

The effect of a factor is defined to be the average change in response produced by a change in the level of that factor. This is frequently called the main effect. For example, consider the data in Table 4.12.

Table 4.12. Data from a 2x2 factorial experiment

		Factor B	
	Level	b_1	b_2
Factor A	a_1	20	30
	a_2	40	52

The main effect of factor A could be thought of as the difference between the average response at the first level of A and the average response at the second level of A. Numerically, this is

$$A = \frac{40+52}{2} - \frac{20+30}{2} = 21$$

That is, increasing factor A from level 1 to level to 2 causes an average increase in the response by 21 units. Similarly, the main effect of B is

$$B = \frac{30+52}{2} - \frac{20+40}{2} = 11$$

If the factors appear at more than two levels, the above procedure must be modified since there are many ways to express the differences between the average responses.

The major advantage of conducting a factorial experiment is the gain in information on interaction between factors. In some experiments, we may find that the difference in response between the levels of one factor is not the same at all levels of the other factors. When this occurs, there is an interaction between the factors. For example, consider the data in Table 4.13.

Table 4.13. Data from a 2x2 factorial experiment

		Factor B	
	Levels	b_1	b_2
Factor A	a_1	20	40
	a_2	50	12

At the first level of factor B, the factor A effect is

A = 50-20 = 30

and at the second level of factor B, the factor A effect is

A = 12-40 = -28

Since the effect of A depends on the level chosen for factor B, we see that there is interaction between A and B.

These ideas may be illustrated graphically. Figure 4.5 plots the response data in Table 4.12. against factor A for both levels of factor B.

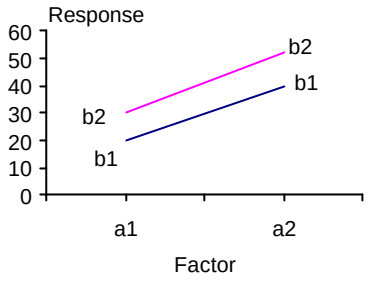


Figure 4.5. Graphical representation of lack of interaction between factors.

Note that the b_1 and b_2 lines are approximately parallel, indicating a lack of interaction between factors A and B.

Similarly, Figure 4.6 plots the response data in Table 4.13. Here we see that the b_1 and b_2 lines are not parallel. This indicates an interaction between factors A and B. Graphs such as these are frequently very useful in interpreting significant interactions and in reporting the results to nonstatistically trained management. However, they should not be utilized as the sole technique of data analysis because their interpretation is subjective and their appearance is often misleading.

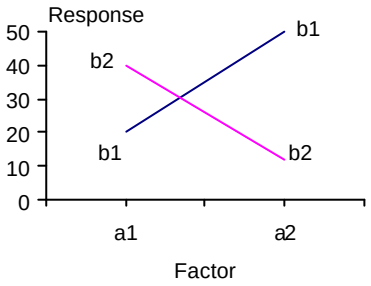


Figure 4.6. Graphical representation of interaction between factors.

Note that when an interaction is large, the corresponding main effects have little practical meaning. For the data of Table 4.13, we would estimate the main effect of A to be

$$A = \frac{50+12}{2} - \frac{20+40}{2} = 1$$

which is very small, and we are tempted to conclude that there is no effect due to A. However, when we examine the effects of A at different levels of factor B, we see that this is not the case. Factor A has an effect, but it depends on the level of factor B *i.e.*, a significant interaction will often mask the significance of main effects. In the presence of significant interaction, the experimenter must usually examine the levels of one factor, say A, with level of the other factors fixed to draw conclusions about the main effect of A.

For most factorial experiments, the number of treatments is usually too large for an efficient use of a complete block design. There are, however, special types of designs developed specifically for large factorial experiments such as confounded designs. Descriptions on the use of such designs can be found in Das and Giri (1980).

4.4.1. Analysis of variance

Any of the complete block designs discussed in sections 4.2 and 4.3 for single-factor experiments is applicable to a factorial experiment. The procedures for randomization and layout of the individual designs are directly applicable by simply ignoring the factor composition of the factorial treatments and considering all the treatments as if they were unrelated. For the analysis of variance, the computations discussed for individual designs are also directly applicable. However, additional computational steps are required to partition the treatment sum of squares into factorial components corresponding to the main effects of individual factors and to their interactions. The procedure for such partitioning is the same for all complete block designs and is, therefore, illustrated for only one case, namely, that of RCBD.

The step-by-step procedure for the analysis of variance of a two-factor experiment on bamboo involving two levels of spacing (Factor A) and three levels of age at planting (Factor A) laid out in RCBD with three replications is illustrated here. The list of the six factorial treatment combinations is shown in Table 4.14, the experimental layout in Figure 4.7, and the data in Table 4.15.

Table 4.14. The 2 x 3 factorial treatment combinations of two levels of spacing and three levels of age.

Age at planting (month)	Spacing (m)	
	10 m x 10 m (a_1)	12 m x 12m (a_2)
6 (b_1)	a_1b_1	a_2b_1
12 (b_2)	a_1b_2	a_2b_2
24 (b_3)	a_1b_3	a_2b_3

Replication I	Replication II	Replication III
a_2b_3	a_2b_3	a_1b_2
a_1b_3	a_1b_2	a_1b_1
a_1b_2	a_1b_3	a_2b_2
a_2b_1	a_2b_1	a_1b_3
a_1b_1	a_2b_2	a_2b_1
a_2b_2	a_1b_1	a_2b_3

Figure 4.7. A sample layout of 2×3 factorial experiment involving two levels of spacing and three levels of age in a RCBD with 3 replications.

Table 4.15. Mean maximum culm height of *Bambusa arundinacea* tested with three age levels and two levels of spacing in a RCBD.

Treatment combination	Maximum culm height of a clump (cm)			Treatment total (T_{ij})
	Rep. I	Rep. II	Rep. III	
a_1b_1	46.50	55.90	78.70	181.10
a_1b_2	49.50	59.50	78.70	187.70
a_1b_3	127.70	134.10	137.10	398.90
a_2b_1	49.30	53.20	65.30	167.80
a_2b_2	65.50	65.00	74.00	204.50
a_2b_3	67.90	112.70	129.00	309.60
Replication total (R_k)	406.40	480.40	562.80	G=1449.60

Step 1. Denote the number of replication by *r*, the number of levels of factor A (*i.e.*, spacing) by *a*, and that of factor B (*i.e.*, age) by *b*. Construct the outline of the analysis of variance as follows:

Table 4.16. Schematic representation of ANOVA of a factorial experiment with two levels of factor A, three levels of factor B and with three replications in RCBD.

Source of variation	Degrees of freedom (*df*)	Sum of squares (*SS*)	Mean square $\left(MS = \frac{SS}{df}\right)$	Computed *F*
Replication	*r*-1	*SSR*	MSR	
Treatment	*ab*- 1	*SST*	MST	$\frac{MST}{MSE}$
A	*a*- 1	*SSA*	MSA	$\frac{MSA}{MSE}$
B	*b*- 1	*SSB*	MSB	$\frac{MSB}{MSE}$
AB	(*a*-1)(*b*-1)	*SSAB*	MSAB	$\frac{MSAB}{MSE}$
Error	(*r*-1)(*ab*-1)	*SSE*	MSE	
Total	*rab* -1	*SSTO*		

Step 2.Compute treatment totals (T_{ij}), replication totals (R_k), and the grand total (G), as shown in Table 4.15 and compute the *SSTO*, *SSR*, *SST* and *SSE* following the procedure described in Section 4.3.3. Let y_{ijk} refer to the observation corresponding to the *i*th level of factor A and *j*th level factor B in the *k*th replication.

$$C.F. = \frac{G^2}{rab} \tag{4.22}$$

$$= \frac{(1449.60)^2}{(3)(2)(3)} = 116741.12$$

$$SSTO = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r} y_{ijk}^2 - C.F. \tag{4.23}$$

$$= \left[(46.50)^2 + (55.90)^2 + \ . \ . \ . \ + (129.00)^2\right] - 116741.12$$
$$= \ 17479.10$$

$$SSR = \frac{\sum_{k=1}^{r} R_k^2}{ab} - C.F. \qquad (4.24)$$
$$= \frac{(406.40)^2 + \ . \ . \ . \ + (562.80)^2}{(2)(3)} - 116741.12$$
$$= 2040.37$$

$$SST = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b} T_{ij}^2}{r} - C.F. \qquad (4.25)$$
$$= \frac{(181.10)^2 + \ . \ . \ . \ + (309.60)^2}{3} - 116741.12$$
$$= 14251.87$$

$$SSE = SSTO - SSR - SST \qquad (4.26)$$
$$= 17479.10 - 2040.37 - 14251.87$$
$$= 1186.86$$

The preliminary analysis of variance is shown in Table 4.17.

Table 4.17. Preliminary analysis of variance for data in Table 4.15.

Source of variation	Degree of freedom	Sum of squares	Mean square	Computed *F*	Tabular *F* 5%
Replication	2	2040.37	1020.187	8.59567*	4.10
Treatment	5	14251.87	2850.373	24.01609*	3.33
Error	10	1186.86	118.686		
Total	17	17479.10			

*Significant at 5% level.

Step 3. Construct the factor A x factor B two-way table of totals, with factor A totals and factor B totals computed. For our example, the Spacing x Age table of totals (*AB*) with Spacing totals (*A*) and Age totals (*B*) computed are shown in Table 4.18.

Table 4.18. The Spacing x Age table of totals for the data in Table 4.15.

Age	Spacing		Total
	a_1	a_2	(B_j)
b_1	181.10	167.80	348.90
b_2	187.70	204.50	392.20
b_3	398.90	309.60	708.50
Total (A_i)	767.70	681.90	G = 1449.60

Step 4. Compute the three factorial components of the treatment sum of squares as:

$$SSA = \frac{\sum_{i=1}^{b} A_i^2}{rb} - C.F. \qquad (4.27)$$

$$= \frac{(767.70)^2 + (681.90)^2}{(3)(3)} - 116741.12$$

$$= 408.98$$

$$SSB = \frac{\sum_{j=1}^{b} B_j^2}{ra} - C.F. \qquad (4.28)$$

$$= \frac{(348.90)^2 + (392.20) + (708.50)^2}{(3)(2)} - 116741.12$$

$$= 12846.26$$

$$SSAB = SST - SSA - SSB \qquad (4.29)$$

$$= 14251.87 - 408.98 - 12846.26$$

$$= 996.62$$

Step 5. Compute the mean square for each source of variation by dividing each sum of squares by its corresponding degrees of freedom and obtain the *F* ratios for each of the three factorial components as per the scheme given in the Table 4.16

Step 6. Enter all values obtained in Steps 3 to 5 in the preliminary analysis of variance of Step 2, as shown in Table 4.19.

Table 4.19. ANOVA of data in Table 4.15 from a 2 x 3 factorial experiment in RCBD.

Source of variation	Degree of freedom	Sum of squares	Mean square	Computed *F*	Tabular *F* 5%
Replication	2	2040.37	1020.187	8.60*	4.10
Treatment	5	14251.87	2850.373	24.07*	3.33
A	1	12846.26	6423.132	3.45	4.96
B	2	408.98	408.980	54.12*	4.10
AB	2	996.62	498.312	4.20*	4.10
Error	10	1186.86	118.686		
Total	17	17479.10			

*Significant at 5% level

Step 7. Compare each of the computed *F* value with the tabular *F* value obtained from Appendix 3, with $f_1 = df$ of the numerator *MS* and $f_2 = df$ of the denominator *MS*, at the desired level of significance. For example, the computed *F* value for

main effect of factor A is compared with the tabular F values (with f_1=1 and f_2=10 degrees of freedom) of 4.96 at the 5% level of significance. The result indicates that the main effect of factor A (spacing) is not significant at the 5% level of significance.

Step 8. Compute the coefficient of variation as:

$$cv = \frac{\sqrt{Error\ MS}}{Grand\ mean} \times 100 \qquad (4.30)$$
$$= \frac{\sqrt{118.686}}{80.53} \times 100 = 13.53\%$$

4.4.2. Comparison of means

In a factorial experiment, comparison of effects are of different types. For example, a 2x3 factorial experiment has four types of means that can be compared.

Type-(1) The two A means, averaged over all three levels of factor B
Type-(2) The three B means, averaged over both levels of factor A
Type (3) The six A means, two means at each of the three levels of factor B
Type (4) The six B means, three means at each of the two levels of factor A

The Type-(1) mean is an average of 3r observations, the Type-(2) is an average of 2r observations and the Type-(3) or Type-(4) is an average of r observations. Thus, the formula $s_{\bar{d}} = \left(2s^2 / r\right)^{1/2}$ is appropriate only for the mean difference involving either Type-(3) or Type-(4) means. For Type-(1) and Type-(2) means, the divisor r in the formula should be replaced by 3r and 2r respectively. That is, to compare two A means averaged over all levels of factor B, the $s_{\bar{d}}$ value is computed as $s_{\bar{d}} = \left(2s^2 / 3r\right)^{1/2}$ and to compare any pair of B means averaged over all levels of factor A, the $s_{\bar{d}}$ value is computed as $\left(2s^2 / 2r\right)^{1/2}$ or simply $\left(s^2 / r\right)^{1/2}$.

As an example, consider the 2x 3 factorial experiment whose data are shown in Table 4.15. The analysis of variance shows a significant interaction between spacing and age, indicating that the effect of age vary with the change in spacing. Hence, comparison between age means averaged over all levels of spacing or between spacing means averaged over all age levels is not useful. The more appropriate mean comparisons are those between age means under the same level of spacing or between spacing means of the same level of age. The comparison between spacing means at the same age level is illustrated in the following. The steps involved in the computation of LSD for comparing two spacing means at same age level are,

Step 1.Compute the standard error of the mean difference following the formula for comparison Type-(3) as

$$s_{\bar{d}} = \sqrt{\frac{2\,\text{Error } MS}{r}} \qquad (4.31)$$

$$= \sqrt{\frac{2(118.686)}{3}} = 8.89 \text{ cm}$$

where the Error *MS* value of 118.686 is obtained from the analysis of variance of Table 4.19.

Step 2. From Appendix 2, obtain the tabular *t* value for error *df* (10 *df*), which is 2.23 at 5% level of significance and compute the LSD as,

$$\text{LSD}_{\alpha} = \left(t_{v;\,\alpha}\right)\left(s_{\bar{d}}\right) = (2.23)(8.89) = 19.82 \text{ cm}$$

Step 3.Construct the Spacing x Age two-way table of means as shown in Table 4.20. For each pair of spacing levels to be compared at the same age level, compute the mean difference and compare it with the LSD value obtained at Step 2. For example, the mean difference in culm height between the two spacing levels at age level of 12 months at planting is 5.6 cm. Because this mean difference is smaller than the LSD value at the 5% level of significance, it is not significant.

Table 4.20. The Spacing x Age table of means of culm height based on data in Table 4.15.

Age at planting (month)	Spacing (m)	
	10 m x 10 m	12 m x 12m
	Mean culm height (cm)	
6	60.37	55.93
12	62.57	68.17
24	132.97	103.20

4.5. Fractional factorial design

In a factorial experiment, as the number of factors to be tested increases, the complete set of factorial treatments may become too large to be tested simultaneously in a single experiment. A logical alternative is an experimental design that allows testing of only a fraction of the total number of treatments. A design uniquely suited for experiments involving large number of factors is the fractional factorial design (FFD). It provides a systematic way of selecting and testing only a fraction of the complete set of factorial treatment combinations. In exchange, however, there is loss of information on some pre-selected effects. Although this information loss may be serious in experiments with one or two factors, such a loss becomes more tolerable with large number of factors. The number of interaction effects increases rapidly with the number of factors involved, which allows flexibility in the choice of the particular effects to be sacrificed. In fact, in cases where some specific effects are known beforehand to be small or unimportant, use of the FFD results in minimal loss of information.

In practice, the effects that are most commonly sacrificed by use of the FFD are high order interactions - the four-factor or five-factor interactions and at times, even the three-factor interaction. In almost all cases, unless the researcher has prior information to indicate otherwise he should select a set of treatments to be tested so that all main effects and two-factor interactions can be estimated. In forestry research, the FFD is to be used in exploratory trials where the main objective is to examine the interactions between factors. For such trials, the most appropriate FFD's are those that sacrifice only those interactions that involve more than two factors.
With the FFD, the number of effects that can be measured decreases rapidly with the reduction in the number of treatments to be tested. Thus, when the number of effects to be measured is large, the number of treatments to be tested, even with the use of FFD, may still be too large. In such cases, further reduction in the size of the experiment can be achieved by reducing the number of replications. Although the use of FFD without replication is uncommon in forestry experiments, when FFD is applied to exploratory trials, the number of replications required can be reduced to the minimum.

Another desirable feature of FFD is that it allows reduced block size by not requiring a block to contain all treatments to be tested. In this way, the homogeneity of experimental units within the same block can be improved. A reduction in block size is, however, accompanied by loss of information in addition to that already lost through the reduction in number of treatments. Although the FFD can thus be tailor-made to fit most factorial experiments, the procedure for doing so is complex and so only a particular class of FFD that is suited for exploratory trials in forestry research is described here. The major features of these selected designs are that they (i) apply only to 2^n factorial experiments where *n*, the number of factors is at least 5, (ii) involve only one half of the complete set of factorial treatment combinations, denoted by 2^{n-1}(iii) allow all main effects and two-factor interactions to be estimated. For more complex plans, reference may made to Das and Giri (1980).

The procedure for layout, and analysis of variance of a 2^{5-1} FFD with a field experiment involving five factors A, B, C, D and E is illustrated in the following. In the designation of the various treatment combinations, the letters a, b, c,…, are used to denote the presence (or high level) of factors A, B, C,… Thus the treatment combination ab in a 2^5 factorial experiment refers to the treatment combination that contains the high level (or presence) of factors A and B and low level (or absence) of factors C, D and E, but this same notation (ab) in a 2^6 factorial experiment would refer to the treatment combination that contains the high level of factors A and B and low level of factors C, D, E, and F. In all cases, the treatment combination that consists of the low level of all factors is denoted by the symbol (1).

4.5.1. Construction of the design and layout

One simple way to arrive at the desired fraction of factorial combinations in a 2^{5-1} FFD is to utilize the finding that in a 2^5 factorial trial, the effect ABCDE can be estimated from the expression arising from the expansion of the term (a-1)(b-1)(c-1)(d-1)(e-1) which is

(a-1)(b-1)(c-1)(d-1)(e-1) = abcde - acde - bcde + cde - abde + ade + bde - de
- abce + ace + bce - ce + abe - ae - be + e
- abcd + acd + bcd - cd + abd - ad - bd + d
+ abc - ac - bc + c - ab + a + b - 1

Based on the signs (positive or negative) attached to the treatments in this expression, two groups of treatments can be formed out of the complete factorial set. Retaining only one set with either negative or positive signs, we get a half fraction of the 2^5 factorial experiment. The two sets of treatments are shown below.

Treatments with negative signs	Treatments with positive signs
acde, bcde, abde, de, abce, ce, ae, be, abcd, cd, ad, bd, ac, bc, ab, 1	abcde, bcde, abde, de, abce, ce, ae, be, abcd, cd, ad, bd, ac, bc, ab, 1

As a consequence of the reduction in number of treatments included in the experiment, we shall not be able to estimate the effect ABCDE using the fractional set. All main effects and two factor interactions can be estimated under the assumption that all three factor and higher order interactions are negligible. The procedure is generalizable in the sense that in a 2^6 experiment, a half fraction can be taken by retaining the treatments with either negative or positive signs in the expansion for (a-1)(b-1)(c-1)(d-1)(e-1)(f-1).

The FFD refers to only a way of selecting treatments with a factorial structure and the resulting factorial combinations can be taken as a set of treatments for the physical experiment to be laid out in any standard design like CRD or RCBD. A sample randomized layout for a 2^{5-1} FFD under RCBD with two replications is shown in Figure 4.8.

1 de	9 ab
2 1	10 adde
3 acde	11 ad
4 ae	12 abce
5 ce	13 be
6 ac	14 bc
7 bcde	15 bcd
8	16

1 abce	9 acde
2 cd	10 bd
3 be	11 de
4 ad	12 bcde
5 ae	13 ce
6 abcd	14 1
7 abce	15 ac
8	16

Figure 4.8. A sample layout of a 2^{5-1} FFD with two replications under RCBD.

4.5.2. Analysis of variance.

The analysis of variance procedure of a 2^{5-1} FFD with 2 replications is illustrated using Yate's method for the computation of sums of squares. This is a method suitable for manual computation of large factorial experiments. Alternatively, the standard rules for the computation of sums of squares in the analysis of variance, by constructing one-way tables of totals for computing main effects, two-way tables of totals for two-factor interactions and so on as illustrated in Section 4.4.1 can also be adopted in this case.

The analysis of 2^{5-1} FFD is illustrated using hypothetical data from a trial whose layout is shown in Figure 4.8 which conforms to that of a RCBD. The response obtained in terms of fodder yield (t/ha) under the different treatment combinations is given in Table 4.21. The five factors were related to different components of a soil management scheme involving application of organic matter, fertilizers, herbicides, water, and lime.

Table 4.21.Fodder yield data from a 2^{5-1} factorial experiment

Treatment combination	Fodder yield (t/ha)		Treatment total (T_i)
	Replication I	Replication II	
acde	1.01	1.04	2.06
bcde	1.01	0.96	1.98
abde	0.97	0.94	1.92
de	0.82	0.75	1.58
abce	0.92	0.95	1.88
ce	0.77	0.75	1.53
ae	0.77	0.77	1.55
be	0.76	0.80	1.57
abcd	0.97	0.99	1.97
cd	0.92	0.88	1.80
ad	0.80	0.87	1.68
bd	0.82	0.80	1.63
ac	0.91	0.87	1.79
bc	0.79	0.76	1.55
ab	0.86	0.87	1.74
1	0.73	0.69	1.42
Replication total (R_j)	13.83	13.69	
Grand total (G)			27.52

The computational steps in the analysis of variance are :

Step1. Construct the outline of the analysis of variance as presented in Table 4.22.

Step 2. Determine the number of real factors (k) each at two levels, whose complete set of factorial treatments is equal to the number of treatments (t) included in the experiment (*i.e.*, $2^k = t$). Then select the specific set of k real factors from the original set of n factors and designate all ($n - k$) factors not included in the set of k as dummy factors. For our example, the $t = 16$ treatment combinations correspond to a complete set of 2^k factorial combinations, with $k = 4$. For simplicity, the first four factors A, B, C and D are designated as the real factors and E as the dummy factor.

Table 4.22. Schematic representation of ANOVA of a 2^{5-1} FFD in RCBD into 2 replications.

Source of variation	Degree of freedom (df)	Sum of squares (SS)	Mean square $\left(MS = \frac{SS}{df}\right)$	Computed F
Block	r-1=1	SSR	MSR	MSR/MSE
A	1	SSA	MSA	MSA/MSE
B	1	SSB	MSB	MSB/MSE
C	1	SSC	MSC	MSC/MSE
D	1	SSD	MSD	MSD/MSE
E	1	$SSE^{@}$	$MSE^{@}$	$MSE^{@}/MSE$
AB	1	$SSAB$	$MSAB$	$MSAB/MSE$
AC	1	$SSAC$	$MSAC$	$MSAC/MSE$
AD	1	$SSAD$	$MSAD$	$MSAD/MSE$
AE	1	$SSAE$	$MSAE$	$MSAE/MSE$
BC	1	$SSBC$	$MSBC$	$MSBC/MSE$
BD	1	$SSBD$	$MSBD$	$MSBD/MSE$
BE	1	$SSBE$	$MSBE$	$MSBE/MSE$
CD	1	$SSCD$	$MSCD$	$MSCD/MSE$
CE	1	$SSCE$	$MSCE$	$MSCE/MSE$
DE	1	$SSDE$	$MSDE$	$MSDE/MSE$

Error	15	*SSE*	*MSE*	
Total	$(r\,2^{5-1})$-1	*SSTO*		

[@] This *SS* is sum of squares due to factor E and is not to be confused with sum of squares due to error (*SSE*) given in the same table later. The error degree of freedom can be obtained by subtracting the degree of freedom for block and the factorial effects from the total degree of freedom.

Step 3. Arrange the t treatments in a systematic order based on the k real factors; Treatments with fewer number of letters are listed first. For example, ab comes before abc, and abc comes before abcd, and so on. Note that if treatment (1) is present in the set of t treatments, it always appears as the first treatment in the sequence. Among treatments with the same number of letters, those involving letters corresponding to factors assigned to the lower-order letters come first. For example, ab comes before ac, ad before bc, and so on. All treatment-identification letters corresponding to the dummy factors are ignored in the arrangement process. For our example, factor E is the dummy factor and, thus, ae is considered simply as a and comes before ab. In this example, the systematic arrangement of the 16 treatments is shown in the first column of Table 4.23. Note that the treatments are listed systematically regardless of their block allocation and the dummy factor E is placed in parenthesis.

Step 4. Compute the t factorial effect totals: Designate the t treatment totals as the initial set or the T_0 values. For our example, the systematically arranged set of 16 T_0 values is listed in the second column of Table 4.23. Next, group the T_0 values into $t/2$ successive pairs. For our example, there are 8 successive pairs : the first pair is 1.42 and 1.54, the second pair is 1.56 and 1.73, and the last pair is 1.97 and 1.96. Add the values of the two treatments in each of the $t/2$ pairs constituted and enter in the second set or the T_1 values. For our example, the first half of the T_1 values are computed as :

2.96 = 1.42 + 1.54
3.29 = 1.56 + 1.73
….
….
3.93 = 1.97 + 1.96

Subtract the first value from the second in each of the $t/2$ pairs constituted under T_0 to constitute the bottom half of the T_1 values. For our example, the second half of the T_1 values are computed as :

-0.12 = 1.42 - 1.54
-0.17 = 1.56 - 1.73
….
….
0.01 = 1.97 - 1.96

The results of these tasks are shown in the third column of Table 4.23.

Reapply tasks done to generate the column T_1, now using the values of T_1 instead of T_0 to derive the third set or the T_2 values. For our example, results of the tasks reapplied to T_1 values to arrive at the T_2 values are shown in the fourth column of Table 4.23. Repeat task (n - 1) times where n is the total number of factors in the experiment. Each time, use the newly derived values of T. For our example, the task is repeated two more times to derive T_3 values and T_4 values as shown in the fifth and sixth columns of Table 4.23.

Table 4.23. Application of Yates' method for the computation of sums of squares of a 2^{5-1} FFD with data in Table 4.21

Treat-ment	T_0	T_1	T_2	T_3	T_4	Factorial Effect Identification		$\frac{(T_4)^2}{r2^{n-1}}$
						Initial	Final	
(1)	1.42	2.96	6.25	12.97	27.52	(G)	(G)	23.667
a(e)	1.54	3.29	6.72	14.55	-1.5	A	AE	0.070
b(e)	1.56	3.30	6.77	-0.87	-0.82	B	BE	0.021
ab	1.73	3.42	7.78	-0.63	0.04	AB	AB	0.000
c(e)	1.52	3.24	-0.29	-0.45	-1.48	C	CE	0.068
ac	1.78	3.53	-0.58	-0.37	0.14	AC	AC	0.001
bc	1.55	3.85	-0.39	0.11	-0.42	BC	BC	0.006
abc(e)	1.87	3.93	-0.24	-0.07	0.44	ABC	D	0.006
d(e)	1.57	-0.12	-0.33	-0.47	-1.58	D	DE	0.078
ad	1.67	-0.17	-0.12	-1.01	-0.24	AD	AD	0.002
bd	1.62	-0.26	-0.29	0.29	-0.08	BD	BD	0.000
abd(e)	1.91	-0.32	-0.08	-0.15	0.18	ABD	C	0.001
cd	1.80	-0.10	0.05	-0.21	0.54	CD	CD	0.009
acd(e)	2.05	-0.29	0.06	-0.21	0.44	ACD	B	0.006
bcd(e)	1.97	-0.25	0.19	-0.01	0.00	BCD	A	0.000
abcd	1.96	0.01	-0.26	0.45	-0.46	ABCD	E	0.007

Step 5. Identify the specific factorial effect that is represented by each of the values of the last set (commonly referred to as the factorial effect totals) derived in Step 4. Use the following guidelines: The first value represents the grand total (G). For the remaining (t - 1) values, assign the preliminary factorial effects according to the letters of the corresponding treatments, with the dummy factors ignored.

For example, the second T_4 value corresponds to treatment combinations a(e) and, hence, it is assigned to the A main effect. The fourth T_4 value corresponds to treatment ab and is assigned to the AB interaction effect, and so on. The results for all 16 treatments are shown in the seventh column of Table 4.23. For treatments involving the dummy factor, adjust the preliminary factorial effects derived as follows. Identify all effects involving the dummy factor that are estimable through the design. For our example, the estimable effects involving

the dummy factor E consist of the main effect of E and all its two-factor interactions AE, BE, CE and DE. Identify the aliases of all effects listed as 'preliminary'. The alias of any effect is defined as its generalized interaction with the defining contrast. The generalized interaction between any two factorial effects is obtained by combining all the letters that appear in the two effects and cancelling all letters that enter twice. For example, the generalized interaction between ABC and AB is AABBC or C. For our example, the defining contrast is ABCDE, the aliases of five effects involving the dummy factor E are: E=ABCD, AE=BCD, BE=ACD, CE=ABD and DE=ABC.

The two factorial effects involved in each pair of aliases (one to the left and another to the right of the equal sign) are not separable (*i.e.*, cannot be estimated separately). For example, for the first pair, E and ABCD, the main effect of factor E cannot be separated from the A BCD interaction effect and, hence, unless one of the pair is known to be absent there is no way to know which of the pairs is the contributor to the estimate obtained.

Replace all preliminary factorial effects that are aliases of the estimable effects involving the dummy factors by the latter. For example, because ABCD (corresponding to the last treatment in Table 4.23) is the alias of E, it is replaced by E. In the same manner, BCDE is replaced by A, ACDE by B and so on. The final results of the factorial effect identification are shown in the eighth column of Table 4.23.

Step 6. Compute an additional column in Table 4.23 as $\frac{(T_4)^2}{r2^{n-1}}$ where r is the number of replications and n is the number of factors in the experiment. The value in this column corresponding to G in the previous column will be the correction factor. The rest of the values in this column will be the sum of squares corresponding to the effects identified in the previous column.

Step 7. Compute the *SS* due to other effects to complete the ANOVA. Let y_{ij} represent the response obtained with the ith treatment in the jth replcation.

$$C.F. = \frac{G^2}{rt} \tag{4.32}$$

$$= \frac{12.37^2}{(2)(16)} = 23.6672$$

$$SSTO = \sum_{i=1}^{t}\sum_{j=1}^{r} y_{ij}^2 - C.F. \tag{4.33}$$

$$= \left[(1.01)^2 + (1.04)^2 + \ \ldots \ + (0.69)^2\right] - 23.6672$$

$$= 0.2866$$

$$SSR = \frac{\sum_{j=1}^{r} R_j^2}{2^{n-1}} - C.F. \qquad (4.34)$$

$$= \frac{(13.83)^2 + (13.69)^2}{2^4} - 23.6672$$

$$= 0.0006$$

$$SST = \frac{\sum_{i=1}^{t} T_i^2}{r} - C.F. \qquad (4.35)$$

$$= \frac{(1.42)^2 + (1.54)^2 + \ldots\ldots\ldots + (1.96)^2}{4} - 23.6672$$

$$= 0.2748$$

$$SSE = SSTO - SSR - SST \qquad (4.36)$$

$$= 0.2866 - 0.2748 - 0.0006$$

$$= 0.01$$

Step 8. Compute the mean square for each source of variation by dividing each *SS* by its *df*. The *MS* corresponding to each factorial effect will be the same as its *SS* in this case because the *df* of such effects is one in each case.

Step 9. Compute the *F* value corresponding to each term in the analysis of variance table by dividing the *MS* values by the error *MS* values. The final analysis of variance is shown in Table 4.24.

Table 4.24. ANOVA of data of Table 4.21 corresponding to a 2^{5-1} factorial experiment.

Source of variation	Degrees of freedom	Sum of squares	Mean square	Computed *F*	Tabular *F* 5%
Replication	1	0.0006	0.0006	0.86^{ns}	4.54
A	1	0.000	0.000	0.00^{ns}	4.54
B	1	0.006	0.006	8.57*	4.54
C	1	0.001	0.001	1.43^{ns}	4.54
D	1	0.006	0.006	8.57*	4.54
E	1	0.007	0.007	10.00*	4.54
AB	1	0.000	0.000	0.00^{ns}	4.54
AC	1	0.001	0.001	1.43^{ns}	4.54
AD	1	0.002	0.002	2.86^{ns}	4.54
AE	1	0.070	0.070	100.00*	4.54
BC	1	0.006	0.006	8.57*	4.54
BD	1	0.000	0.000	0.00^{ns}	4.54
BE	1	0.021	0.021	30.00*	4.54

CD	1	0.009	0.009	12.86*	4.54
CE	1	0.068	0.068	97.14*	4.54
DE	1	0.078	0.078	111.43*	4.54
Error	15	0.010	0.0007		
Total	31	0.2866			

* Significant at 5% level, ^{ns} = nonsignificant at 5% level

Step 11. Compare each computed *F* value with the corresponding tabular *F* values, from Appendix 3, with f_1 = *df* of the numerator *MS* and f_2 = error *df*. The results show that main effects B, D and E and the two factor interactions AE, BC, BE, CD, CE and AE are highly significant and main effects A and C and the two factor interactions AB, AC, AD and BD are non significant.

4.5.3. Comparison of means

The procedure described in section 4.4.2 for comparison of means in the case of complete factorial experiments is applicable to the case of FFD as well but remembering the fact that means of only up to two-way tables can only be compared using the multiple comparison procedure in the case of 2^{5-1} factorial experiment.

4.6. Split plot design

The split plot design is specifically suited for a two-factor experiment wherein levels of one of the factors require large plot size for execution and also show large differences in their effects. In such a situation, the experiment will consist of a set of large plots called *main plots* in which levels for the *main plot factor* are assigned. Each main plot is divided into *subplots* to which the second factor, called the *subplot factor*, is assigned. Thus, each main plot becomes a block for the subplot treatments (*i.e.*, the levels of the subplot factor). The assignment of the main plot factor can, in fact, follow any of the patterns like the completely randomized design, randomized complete block, or latin square design but here, only the randomized complete block is considered for the main plot factor because it is perhaps the most appropriate and the most commonly used design for forestry experiments.

With a split plot design, the precision for the measurement of the effects of the main plot factor is sacrificed to improve that of the subplot factor. Measurement of the main effect of the subplot factor and its interaction with the main plot factor is more precise than that obtainable with a randomized complete block design. On the other hand, the measurement of the effects of the main plot treatments (*i.e.*, the levels of the main plot factor) is less precise than that obtainable with a randomized complete block design.

4.6.1. Layout

There are two separate randomization processes in a split plot design-one for the main plots and another for the subplots. In each replication, main plot treatments are first

randomly assigned to the main plots followed by a random assignment of the subplot treatments within each main plot.

For illustration, a two-factor experiment involving four levels of nitrogen (main plot treatments) and three eucalyptus clones (subplot treatments) in three replications is used. Here, fertilzer levels were chosen for the main plots mainly for the convenience in its application, easiness in controlling the leaching effect and to detect the presence of possible interaction between fertilizers and the clones. The steps in the randomization and layout of a split plot design are shown, using a as the number of main plot treatments, b as the number of subplot treatments, and r as the number of replications.

Step 1. Divide the experimental area into $r = 3$ blocks, each of which is further divided into $a = 4$ main plots, as shown in Figure 4.9.

Step 2. Following the RCBD randomization procedure with $a = 4$ treatments and $r = 3$ replications randomly assign the 4 nitrogen treatments to the 4 main plots in each of the 3 blocks. The result may be as shown in Figure 4.10.

Step 3. Divide each of the $ra = 12$ main plots into $b = 3$ subplots and following the RCBD randomization procedure for $b = 3$ treatments and $ra = 12$ replications, randomly assign the 3 clones to the 3 subplots in each of the 12 main plots. The result may be as shown in Figure 4.11.

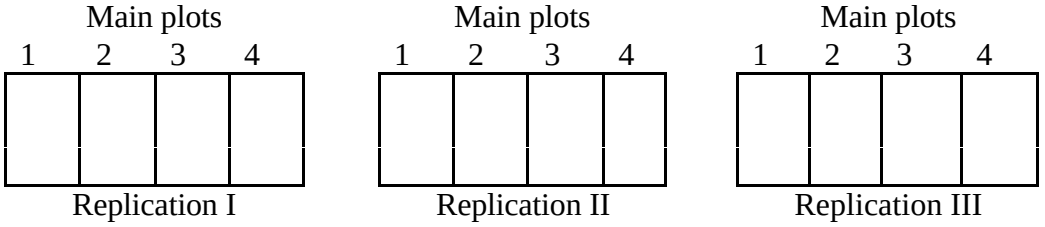


Figure 4.9. Division of the experimental area into three blocks (replications) each consisting of four main plots, as the first step in laying out of a split plot experiment involving three replications and four main plot treatments.

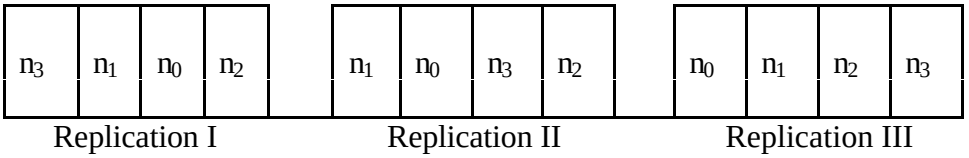
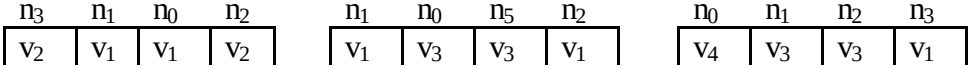


Figure 4.10. Random assignment of four nitrogen levels (n_0, n_1, n_2 and n_3) to the four main plots in each of the three replications of Figure 4.9.



v_1	v_3	v_2	v_3
v_3	v_2	v_3	v_1

Replication I

v_3	v_1	v_2	v_2
v_2	v_2	v_1	v_3

Replication II

v_2	v_4	v_2	v_3
v_1	v_1	v_4	v_2

Replication III

Figure 4.11. A sample layout of a split plot design involving three eucalyptus clones(v_1, v_2 and v_3) as subplot treatments and four nitrogen levels (n_0, n_1, n_2 and n_3) as main plot treatments, in three replications.

Note that the field layout of a split plot design as illustrated by Figure 4.11 has the following important features: (i) The size of the main plot is *b* times the size of the subplot. In our example with 3 varieties ($b = 3$) the size of the main plot is 3 times the subplot size (ii) Each main plot treatment is tested *r* times whereas each subplot treatment is tested *ar* times. Thus, the number of times a subplot treatment is tested will always be larger than that for the main plot and is the primary reason for more precision for the subplot treatments relative to the main plot treatments. In our example, each of the 4 levels of nitrogen is tested 3 times but each of the 3 clones is tested 12 times.

4.6.2. Analysis of variance

The analysis of variance of a split plot design is divided into the *main plot analysis* and the *subplot analysis*. The computations are shown with the data from a two-factor experiment in eucalyptus involving two silvicultural treatments (pit size) and four fertiliser treatments. The data on height of plants after one year of planting are shown in Table 4.25.

Table 4.25. Data on height (cm) of *Eucalyptus tereticornis* plants from a field trial under split plot design.

Fertiliser	Height (cm)		
	Replication I	Replication II	Replication III
Pit size (30 cm x 30 cm x 30 cm) - p_0			
f_0	25.38	61.35	37.00
f_1	46.56	66.73	28.00
f_2	66.22	35.70	35.70
f_3	30.68	58.96	21.58
Pit size (40 cm x 40 cm x 40 cm) - p_1			
f_0	19.26	55.80	57.60
f_1	19.96	33.96	31.70
f_2	22.22	58.40	51.98
f_3	16.82	45.60	26.55

Let A denote the main-plot factor (pit size) and B, the subplot factor (fertiliser treatments). Carry out the analysis of variance as follows:

Step 1. Construct an outline of the analysis of variance for a split plot design as follows.

Table 4.26. Schematic representation of ANOVA of a split plot experiment.

Source of variation	Degree of freedom (df)	Sum of squares (SS)	Mean square $\left(MS = \frac{SS}{df}\right)$	Computed F
Replication	r - 1	SSR	MSR	MSR/MSE_a
A	a - 1	SSA	MSA	MSR/MSE_a
Error (a)	$(r - 1)(a - 1)$	SSE_a	MSE_a	
B	b - 1	SSB	MSB	MSR/MSE_b
AB	$(a - 1)(b - 1)$	$SSAB$	$MSAB$	MSR/MSE_b
Error (b)	$a(r - 1)(b - 1)$	SSE_b	MSE_b	
Total	rab - 1	$SSTO$		

Step 2. Construct two tables of totals as:

(i) The replication x factor A two-way table of totals, with the replication totals, Factor A totals and grand total : For our example, the replication x pit size table of totals ($(RA)_{ki}$), with the replication totals (R_k), pit size totals (A_i) and the grand total (G) computed, is shown in Table 4.27.

Table 4.27. The replication x pit size table of height totals computed from data in Table 4.25.

Pit size	Rep. I	Rep. II	Rep. III	(A_i)
p_0	168.84	222.74	122.28	513.86
p_1	78.26	193.76	167.83	439.85
Rep. total (R_k)	247.10	416.50	290.10	
Grand total (G)				953.70

(ii) The factor A x factor B two-way table of totals, with factor B totals : For our example, the pit size x fertilizer treatment table of totals (AB), with the fertilizer treatment totals (B_j) computed, is shown in Table 4.28.

Table 4.28. The pit size x fertilizer treatment table of height totals computed from data in Table 4.25

Pit size	Fertilizer treatment			
	f_0	f_1	f_2	f_3
p_0	123.73	141.29	137.62	111.22
p_1	132.66	85.62	132.60	88.97
Total (B_j)	256.39	226.91	270.22	200.19

Step 3. Compute the correction factor and sums of squares for the main plot analysis as follows. Let y_{ijk} refer to the response observed ith main plot, jth subplot in the rth replication.

$$C.F. = \frac{G^2}{rab} \tag{4.37}$$
$$= \frac{(953.70)^2}{(3)(2)(4)} = 37897.92$$

$$SSTO = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r} y_{ijk}^2 - C.F. \tag{4.38}$$
$$= [(25.38)^2 + (46.56)^2 + \ldots + (26.55)^2] - 37897.92$$
$$= 6133.10$$

$$SSR = \frac{\sum_{k=1}^{r} R_k^2}{ab} - C.F. \tag{4.39}$$
$$= \frac{(247.10)^2 + (416.50)^2 + (290.10)^2}{(2)(4)} - 37897.92$$
$$= 1938.51$$

$$SSA = \frac{\sum_{i=1}^{a} A_i^2}{rb} - C.F. \tag{4.40}$$
$$= \frac{(513.86)^2 + (439.85)^2}{(3)(4)} - 37897.92$$
$$= 228.25$$

$$SSE_a = \frac{\sum_{k=}^{r}\sum_{i=1}^{a}\left((RA)_{ki}\right)^2}{b} - C.F - SSR - SSA \tag{4.41}$$
$$= \frac{(168.84)^2 + \ldots + (167.83)^2}{(4)} - 40064.68$$

= 1161.70

Step 4. Compute the sums of squares for the subplot analysis as:

$$SSB = \frac{\sum_{j=1}^{b} B_j^2}{ra} - C.F. \qquad (4.42)$$

$$= \frac{(256.39)^2 + \ldots + (200.19)^2}{(3)(2)} - 37897.92$$

$$= 488.03$$

$$SSAB = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\left((AB)_{ij}\right)^2}{r} - C.F. - SSB - SSA \qquad (4.43)$$

$$= \frac{(123.73)^2 + \ldots + (88.97)^2}{3} - 37897.92 - 488.03 - 1161.70$$

$$= 388.31$$

$$SSE_b = SSTO - SSR - SSA - SSB - SSAB - SSE_a \qquad (4.44)$$

$$= 6133.10 - 1938.51 - 228.25 - 488.03 - 388.31$$

$$= 3090.00$$

Step 5. For each source of variation, compute the mean square by dividing the *SS* by its corresponding *df*. The *F* value for each effect that needs to be tested is to be computed by dividing each mean square by the corresponding error term as shown in Table 4.26.

Step 6. Enter all values obtained from Steps 3 to 5 in the ANOVA table as shown in Table 4.29; and compare each of the computed *F* values with its corresponding tabular *F* values and indicate its significance or otherwise by the appropriate asterisk notation. For each effect whose computed *F* value is not less than 1, obtain the corresponding tabular *F* value, from Appendix 3, with $f_1 = df$ of the numerator *MS* and $f_2 = df$ of the denominator *MS*, at the prescribed level of significance. For example, the tabular *F* value for testing the AB effect is 3.49 at 5% level of significance for 3 and 12 degrees of freedom.

Table 4.29. ANOVA of data in Table 4.20 from a split plot design

Source of variation	Degree of freedom	Sum of squares	Mean square	Computed *F*	Tabular *F* 5%
Replication	2	1938.51	969.26		
A	1	228.25	228.25	0.3930^{ns}	4.75
Error (a)	2	1161.70	580.85		
B	3	488.03	162.68	0.6318^{ns}	3.49

AB	3	388.31	129.44	0.5027^{ns}	3.49
Error (b)	12	3090.00	257.50		
Total	23	37897.92			

^{ns} Nonsignificant at 5% level

Step 7. Compute the two coefficients of variation, one corresponding to the main plot analysis and another corresponding to the subplot analysis.

$$cv(\text{a}) = \frac{\sqrt{\text{Error(a)}\,MS}}{\text{Grand mean}} \text{ x } 100 \qquad (4.45)$$

$$= \frac{\sqrt{228.25}}{39.54} \text{ x } 100 = 60.95\%$$

$$cv(\text{b}) = \frac{\sqrt{\text{Error(b)}\,MS}}{\text{Grand mean}} \text{ x } 100 \qquad (4.46)$$

$$= \frac{\sqrt{257.50}}{39.54} \text{ x } 100 = 40.58\%$$

The value of *cv*(a) indicates the degree of precision attached to the main plot factor. The value of *cv*(b) indicates the precision of the subplot factor and its interaction with the main-plot factor. The value of *cv*(b) is expected to be smaller than that of *cv*(a) because, as indicated earlier, the factor assigned to the main plot is expected to be measured with less precision than that assigned to the subplot. In our example, the value of *cv*(b) is smaller than that of *cv*(a) but both of them were high enough to mask any possible treatment differences turning the all the factor effects in the ANOVA nosignificant.

4.6.3. Comparison of treatments

In a split plot design, there are four different types of pair comparisons. Each requires its own set of LSD values. These comparisons are:

Type-(1). Comparisons between two main plot treatment means averaged over all subplot treatments.

Type-(2). Comparison between two subplot treatment means averaged over all main plot treatments.

Type-(3). Comparison between two subplot treatment means at the same main plot treatment.

Type-(4). Comparison between two main plot treatment means at the same or different subplot treatments (*i.e.*, means of any two treatment combinations).

Table 4.30 gives the formula for computing the appropriate standard error of the mean difference ($s_{\bar{d}}$) for each of these types of pair comparison.

Table 4.30. Standard error of the mean difference for each of the four types of pair comparison in a split plot design.

Type of pair comparison		$s_{\bar{d}}$
Type-(1) :	Between two main plot means (averaged over all subplot treatments)	$\sqrt{\frac{2E_a}{rb}}$
Type-(2) :	Between two subplot means (averaged over all main plot treatments)	$\sqrt{\frac{2E_b}{ra}}$
Type-(3) :	Between two subplot means at the same main plot treatment	$\sqrt{\frac{2E_b}{r}}$
Type-(4) :	Between two main plot means at the same or different subplot treatments	$\sqrt{\frac{2[(b-1)E_b + E_a]}{rb}}$

Note : $E_a = MSE_a$, $E_b = MSE_b$, r = number of replications, a = number of main plot treatments, and b = number of subplot treatments.

When the computation of $s_{\bar{d}}$ involves more than one error term, such as in comparisons of Type-(4), the tabular t values from Appendix 2 cannot be used directly and weighted tabular t values are to be computed. The formula for weighted tabular t values in such a case is given below.

$$\text{Weighted tabular } t \text{ value} = \frac{(b-1)\,E_b t_b + E_a t_a}{(b-1)\,E_b + E_a} \qquad (4.47)$$

where t_a is the t value for Error (a) df and t_b is the t value for Error (b) df.

As an example, consider the 2 x 4 factorial experiment whose data are shown in Table 4.25. Although the analysis of variance (Table 4.29) shows all the three effects (the two main effects and the interaction effect) as nonsignificant, for the purpose of illustration, consider the case where there is a significant interaction between pit size and fertiliser indicating the variation in fertilizer effect with the changing pit size. In such a case, comparison between the means of pit size levels pooled over all fertilizer levels or that between fertilizer levels averaged over all levels of pit size will not be valid. The more appropriate comparisons will be those between fertilizer means under the same pit size levels or between pit size means at same fertilizer level. Thus the steps involved in the computation of the LSD for comparing two subplot means at the same main plot treatment are:

Step 1.Compute the standard error of the difference between means following the formula for Type-(3) comparison of Table 4.30.

$$s_{\bar{d}} = \sqrt{\frac{2E_b}{r}}$$
$$= \sqrt{\frac{2(257.5)}{3}} = 3.27$$

Step 2. Following the formula $\text{LSD}_{\alpha} = \left(t_{v;\,\alpha}\right)\left(s_{\bar{d}}\right)$ compute the LSD value at 5% level of significance using the tabular *t* value with 12 degrees of freedom of Error(b).

$$\text{LSD}_{.05} = (2.18)(3.27) = 7.129$$

Step 3.Construct the pit size x fertilizer two-way table of mean differences in height as shown in Table 4.31. Compare the observed differences in the mean height among the fertilzer levels at each pit size with the LSD value computed at Step 2 and identify significant differences if any.

Table 4.31. Difference between mean height of eucalyptus plants with four fertilizer levels at the pit size of 30 cm x 30 cm x 30 cm based on the data in Table 4.25.

	Difference in mean height (cm) at p_0			
	f_0	f_1	f_2	f_3
f_0	0.00	-5.86	-4.63	4.17
f_1		0.00	1.23	10.03
f_2			0.00	8.80
f_3				0.00
	Difference in mean height (cm) at p_1			
	f_0	f_1	f_2	f_3
f_0	0.00	15.68	0.02	14.56
f_1		0.00	-15.66	-1.12
f_2			0.00	14.54
f_3				0.00

4.7. Lattice designs

Theoretically, the complete block designs, such as RCBD are applicable to experiments with any number of treatments. However, these complete block designs become inefficient as the number of treatments increases, because the blocks loose their homogeneity due to the large size. An alternative class of designs for single-factor experiments having a large number of treatments is the incomplete block designs. As the name implies, each block in an incomplete block design does not contain all treatments and so a reasonably small block size can be maintained even if the number of treatments is large. One consequence of having incomplete blocks in the design is

that the comparison of treatments appearing together in a block will be made with greater precision than those not doing so. This difficulty can be circumvented by seeing that in the overall design, every pair of treatments appear together in some block or other equal number of times. Such designs are called balanced designs. Since it calls for large number of replications to achieve complete balance, one may go for partially balanced designs wherein different levels of precision for comparison are admitted for different groups of treatments. One commonly used class of incomplete block designs in forestry experiments is the class of lattice designs wherein the number of treatments is a perfect square and the blocks can be grouped into complete sets of replications. A special case of lattice designs is the simple lattice designs which is discussed in the following.

4.7.1. Simple lattice design

Simple lattice design is also called double lattice or square lattice design. As the number of treatments is to be a perfect square, the design may be constructed for number of treatments such as 9, 16, 25, 36, 49, 64, 81, 121 etc. This design needs two replications and is only a partially balanced design in the sense that the treatments form two groups and the degree of precision of treatment comparison differs between these two groups. The construction and layout of the design are exemplified for the case of 25 treatments.

Step 1. Assign numbers 1 to 25 to the treatments at random. This is necessary to avoid any kind of bias of unknown origin affecting treatment effects.

Step 2. Arrange the treatment numbers from 1 to 25 in the form of a square as given in Figure 4.12.

1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	25

Figure 4.12. Initial arrangement of treatments in simple lattice design

Step 3. Group the treatments by rows. This results in the groupings (1, 2, 3, 4, 5), (6, 7, 8, 9, 10), (11, 12, 13, 14, 15), (16, 17, 18, 19, 20) and (21, 22, 23, 24, 25). Each block will now constitute a group of treatments for one block and five such blocks constitute one complete replication. This special way of row-wise grouping is generally known as X-grouping or A-grouping.

Step 4. Again group the treatments by columns. This results in the groupings (1, 6, 11, 16, 21), (2, 7, 12, 17, 22), (3, 8, 13, 18, 23), (4, 9, 14, 19, 24) and (5, 10, 15, 20, 25). Each group will now constitute a group of treatments for one block and five such blocks will constitute one complete replication. This

special way of column-wise grouping is generally known as Y-grouping or B-grouping.

The X-grouping and Y-grouping ensure that treatments that have occurred together in the same block once do not appear together in the same block again. The above two sets of groupings will appear as in Figure 4.13 before randomization is done.

Replication I (X- grouping)

Block No. 1	1	2	3	4	5
Block No. 2	6	7	8	9	10
Block No. 3	11	12	13	14	15
Block No. 4	16	17	18	19	20
Block No. 5	21	22	23	24	25

Replication II (Y-grouping)

Block No.6	1	6	11	16	21
Block No.7	2	7	12	17	22
Block No.8	3	8	13	18	23
Block No.9	4	9	14	19	24
Block No.10	5	10	15	20	25

Figure 4.13. Two replications of a simple lattice design before randomizaion

Step 5. Within each replication, the treatment groups are allocated at random to the different blocks. This randomization is to be done separately for each replication. The allocation of treatments to plots within each block should also be randomized. The randomization should be done separately for each group independently for each replication. Finally while laying down the replications in the field, the X and Y replications should themselves be randomized over the field locations. This procedure of treatment and replication allocations ensures elimination of any kind of unknown

systematic variations affecting the treatment effects. As a result of complete randomization the actual layout plan might appear as shown in Figure 4.14.

Block No. 5	25	24	21	23	22
Block No. 4	20	19	18	17	16
Block No. 1	5	4	1	3	2
Block No. 3	13	14	15	12	11
Block No. 2	6	9	7	10	8
Block No. 6	16	6	1	21	11
Block No. 9	19	4	9	14	24
Block No. 7	7	2	17	22	12
Block No. 10	5	20	25	10	15
Block No. 8	23	3	8	18	13

Figure 4.14. Randomized layout plan of simple lattice design.

If the blocks within each replication are contiguous, under certain conditions, this will allow the analysis of the whole experiment as a RCBD. It was mentioned already that simple lattice design requires a minimum of two replications, one with X-grouping and other with Y-grouping of the treatments. If more than two replications are desired for this design, then it should be in multiples of two only, as both groups (X and Y) will have to be repeated equal number of times. The procedures of treatment allocations remain the same as above.

4.7.2. Analysis of variance for a simple lattice design

The steps involved in the analysis of variance when the basic design of simple lattice is repeated only once, are indicated below along with the checks on the computations where they are found important. The material drawn for this illustration is from an experiment conducted at Vallakkadavu, in Kerala involving 25 clones of *Eucalyptus grandis.*

Table 4.32 below shows the actual field layout showing positions of the blocks and allocation of treatments within each block under randomized condition. The value in the top left corner of each cell is the clone number and the value in bottom right corner shows the mean height of trees in the plot, one year after planting. Unlike in the case of complete block designs, the analysis of variance for incomplete complete block designs involves adjustments on sums of squares for treatments and blocks because of the incomplete structure of the blocks.

Table 4.32. Layout plan for 5 x 5 double lattice showing the height growth (cm) of *Eucalyptus grandis* clones.

Replication - I

Block No. 5	25 96.40	24 107.90	21 119.30	23 134.30	22 129.20
Block No. 4	20 148.00	19 99.20	18 101.40	17 98.00	16 106.70
Block No. 1	5 158.00	4 122.50	1 136.70	3 123.60	2 113.50
Block No. 3	13 126.80	14 101.60	15 111.70	12 117.30	11 108.20
Block No. 2	6 126.80	9 127.00	7 119.10	10 90.90	8 130.40

Replication - II

Block No. 6	16 169.60	6 157.90	1 124.10	21 134.50	11 112.10
Block No. 9	19 110.30	4 153.40	9 87.10	14 95.30	24 120.50
Block No. 7	7 125.60	2 151.10	17 115.90	22 168.40	12 93.30
Block No. 10	5 126.00	20 106.80	25 137.60	10 132.90	15 117.30

Block No. 8	23	3	8	18	13
	133.10	142.70	115.80	128.90	115.80

Step 1. Arrange the blocks within each group (X and Y groups) and treatments within each block systematically along with the observations as in Table 4.33.

Table 4.33. Systematic arrangement of blocks and the treatments within the blocks of Table 4.32.

Replication - I (X-group)

Block No. 1	1	2	3	4	5
	136.70	113.50	123.60	122.50	158.00
Block No. 2	6	7	8	9	10
	126.80	119.10	130.40	127.00	90.90
Block No. 3	11	12	13	14	15
	108.20	117.30	126.80	101.60	111.70
Block No. 4	16	17	18	19	20
	106.70	98.00	101.40	99.20	148.00
Block No. 5	21	22	23	24	25
	119.30	129.20	134.30	107.90	96.40

Replication - II (Y-group)

Block No. 6	1	6	11	16	21
	124.10	157.90	112.10	169.60	134.50
Block No. 7	2	7	12	17	22
	151.10	125.60	93.30	115.90	168.40
Block No. 8	3	8	13	18	23
	142.70	115.80	115.80	128.90	133.10
Block No. 9	4	9	14	19	24
	153.40	87.10	95.30	110.30	120.50
Block No. 10	5	10	15	20	25
	126.00	132.90	117.30	106.80	137.60

Step 2. Set up a table of treatment totals by summing up the yields for each clone from both replications. This is shown in Table 4.34. These totals are not adjusted for any block effects.

Table 4.34. Treatment (clone) totals

1	2	3	4	5
260.80	264.60	266.30	275.90	284.00
6	7	8	9	10

284.70	244.70	246.20	214.10	223.80
11 220.30	12 210.60	13 242.60	14 196.90	15 229.00
16 276.30	17 213.90	18 230.30	19 209.50	20 254.80
21 253.80	22 297.60	23 267.40	24 228.40	25 234.00

Step 3. Compute the block totals $B_1, B_2, \ldots, B_{10}$ for all the blocks by summing the observations occurring in each block. For example, the block total B_1 for the first block is given by

$B_1 = 136.70+113.50+123.60+122.50+158.00 = 654.30$

Compute the total for each replication by summing the block totals within each replication. For Replication I,

$$R_1 = B_1 + B_2 + B_3 + B_4 + B_5 \quad (4.48)$$
$$= 654.30 + 594.20 + 565.60 + 553.30 + 587.10$$
$$= 2954.50$$

Compute the grand total as $G = R_1 + R_2$ (4.49)
$$= 2954.50 + 3176.00$$
$$= 6130.50$$

Step 4. Construct an outline for the ANOVA table of simple lattice design.

Table 4.35. Schematic representation of ANOVA table of simple lattice design

Source of variation	Degrees of freedom (df)	Sum of squares (SS)	Mean square $\left(MS = \frac{SS}{df}\right)$	Computed F
Replication	r - 1	SSR	MSR	$\frac{MSR}{MSE}$
Treatment (unadj.)	k^2 - 1	SST (unadj.)	MST (unadj.)	$\frac{MST\text{ (unadj.)}}{MSE}$
Blocks within replication (adj.)	$r(k\text{-}1)$	SSB (adj.)	MSB (adj.)	$\frac{MSB\text{ (adj.)}}{MSE}$
Intra-block error	$(k\text{-}1)(rk\text{-}k\text{-}1)$	SSE	MSE	
Total	rk^2 - 1	$SSTO$		

Step 5. Obtain the total sum of squares, replication sum of squares and unadjusted treatment sum of squares. For this, first compute the correction factor ($C.F.$).

$$C.\,F. = \frac{G^2}{n} \qquad (4.50)$$

where $n = rk^2$

r = Number of replications
k^2 = Number of treatments
k = Number of plots in a block

$$C.\,F. = \frac{(6130.50)^2}{2 \times 25} = 751660.61$$

For total sum of squares, find the sum of squares of all the observations in the experiment and subtract the correction factor.

$$SSTO = \sum y^2 - C.F. \qquad (4.51)$$

$$= \{ (136.70)^2 + (113.50)^2 + \ldots\ldots + (137.60)^2 \} - C.\,F.$$
$$= 770626.43 - 751660.61 = 18965.83$$

Compute the replication sum of squares as

$$SSR = \frac{R_1^2 + R_2^2}{k^2} - C.F. \qquad (4.52)$$

$$= \frac{(2\,954.50\ )^2 + (3176.00)^2}{25} - 751660.61$$
$$= 752641.85 - 751660.61 = 981.245$$

Compute unadjusted treatment sum of squares as

$$SST\ (\text{unadj.}) = \sum_{i=1}^{t} \frac{T_i^2}{r} - C.\,F. \qquad (4.53)$$

$$= \frac{(260.80)^2 + (264.60)^2 + \ldots + (234.00)^2}{2} - 751660.61$$
$$= 760747.90 - 751660.61 = 9087.29$$

Step 6. Compute for each block, in Replication 1 (X-group) an adjusted block total C_b subtracting each block total in Replication I, from the corresponding column total in Replication II (Y-group) containing the same set of varieties as in blocks of Replication I as shown in Table 4.36. Similarly, compute for each block total in Replication II, an adjusted block total by subtracting each block total in Replication II, from corresponding column total in Replication I (X-group) containing the same set of varieties as in blocks of Replication II as shown in Table 4.37. Obtain the total of C_b values for each replication and check if they add up to zero.

Total of C_b values for Replication I = U_1 = 221.50
Total of C_b values for Replication II = U_2 = -221.50

This check ensures the arithmetical accuracy of the calculation in the previous steps.

Table 4.36. Computation of C_b values for blocks in Replication I

Block	Replication II Column total	Replication I Block total	C_b - value
1	697.30	654.30	43.00 (C_1)
2	619.30	594.20	25.10 (C_2)
3	533.80	565.60	-31.80 (C_3)
4	631.50	553.30	78.20 (C_4)
5	694.10	587.10	107.00 (C_5)
Total	3176.00	2954.50	221.50 (R_{C1})

Table 4.37. Computation of C_b values for blocks in Replication II

Block	Replication I Column total	Replication II Block total	C_b- value
6	597.70	698.20	-100.50 (C6)
7	577.10	654.30	-77.20 (C7)
8	616.50	636.30	-19.80 (C8)
9	558.20	566.60	-8.40 (C9)
10	605.00	620.60	-15.60 (C10)
Total	2954.50	3176.00	-221.50 (R_{C2})

The adjusted block sum of squares is then given by :

$$SSB\,(\text{adj.}) = \frac{\sum_{b=1}^{10} C_b^2}{kr(r-1)} - \frac{\sum_{j=1}^{2} R_{Cj}^2}{k^2 r(r-1)} \qquad (4.54)$$

where r = Number of replications,

k = Number of treatments per block.

$$SSB\,(\text{adj.}) = \frac{(-43.00)^2 + \ldots\ldots + (-15.60)^2}{(2)(5)(1)} - \frac{(221.50)^2 + (-221.50)^2}{(5^2)(2)(1)}$$

$$= 3782.05 - 1962.49 = 1819.56$$

Finally, the error sum of squares is obtained by subtraction.

$$SSE = SSTO - SSR - SST\,(\text{unadj.}) - SSB\,(\text{adj.}) \qquad (4.55)$$

$$= 18965.83 - 981.24 - 9087.29 - 1819.56$$

$$= 7077.73$$

Note that the sum of squares due to error (*SSE*) computed here indicates that part of the variation (in the response variable) between plots within each block caused by uncontrolled extraneous factors and is generally known as the intra-block error. The adjusted block sum of squares is called inter block error.

Step 7.After obtaining the different sums of squares, enter all the values in the ANOVA table as shown as Table 4.38. The mean squares are obtained by dividing the sum of squares by degrees of freedom as usual.

Table 4.38. ANOVA table of simple lattice design using data in Table 4.32.

Source of variation	Degrees of freedom (*df*)	Sum of squares (*SS*)	Mean square $\left(MS = \frac{SS}{df}\right)$	Computed *F*
Replication	1	981.24	981.24	2.218
Treatment (unadj.)	24	9087.29	378.64	0.856
Block within replication (adj.)	8	1819.56	227.44	0.514
Intra-block error	16	7077.73	442.36	
Total	49	18965.83		

The treatment mean square, as presented in the ANOVA table (Table 4.38), is not adjusted for block effects. As pointed out earlier, the treatment means are not free from block effects. As a result, the ANOVA does not provide a valid *F*-test for testing the treatment differences. Before applying *F*-test, the treatment means are to be adjusted for block effects and adjusted sum of squares for treatments will have to be computed. The procedure for this is given in Step 9. Though this procedure may be adopted when necessary, it is not always necessary to go through further computation unless it is indicated. For instance, in field trial with a large number of treatments, a significant difference among treatment means may generally be expected. As a less sensitive test for treatment difference, a preliminary RCBD analysis may be carried out from the results of Table 4.38.

Step 8. Preliminary RCBD analysis : The error sum of squares for RCBD analysis is obtained by first pooling the inter block error with intra-block error and completing the ANOVA table as follows.

Pooled error = Inter block error + Intra-block error (4.56)
= 1819.56 + 7077.73
= 8897.29

Table 4.39. ANOVA table for preliminary RCBD analysis.

Source of variation	Degrees of freedom (*df*)	Sum of squares (*SS*)	Mean square $\left(MS = \frac{SS}{df}\right)$	Computed *F*
Replication	1	981.24	981.24	
Treatment	24	9087.29	378.64	1.02
Pooled error	24	8897.29	370.72	
Total	49	18965.83		

The observed F-value of 1.02 obtained as the ratio of treatment mean square and pooled error mean square is less than the table value of F at which was 1.98 at 5% level of significance for (24, 24) degrees of freedom, and is suggestive of the fact that the treatments do not differ significantly at 5% level of significance. Since this preliminary RCBD analysis has ended up with a nonsignificant value for F, there is a need to employ a more appropriate F-test by adjusting the treatment sum of squares for block effects, since such a procedure would only increase the sensitivity by which the testing is made but not decrease it. The procedure for effecting such an adjustment to the treatment sum of squares for obtaining a more appropriate F-test for testing treatment differences is given in Step 9.

Step 9. Computation of treatment sum of squares adjusted for block effects : First, obtain the unadjusted block sum of squares within replications. Since we have already calculated the block sums $B_1, B_2, \ldots, B_{10}$ in Step 3, this is easily computed as follows :

Unadjusted block SS for Replication I = SSB_1(unadj.)

$$= \frac{B_1^2 + B_2^2 + \ldots + B_5^2}{k} - \frac{R_1^2}{k^2} \qquad (4.57)$$

$$= \frac{(654.30)^2 + \ldots + (587.10)^2}{5} - \frac{(2954.50)^2}{25}$$

$$= 1219.75$$

Unadjusted block SS for Replication II = SSB_2 (unadj.)

$$= \frac{B_6^2 + B_7^2 + \ldots + B_{10}^2}{k} - \frac{R_2^2}{k^2} \qquad (4.58)$$

$$= \frac{(698.20)^2 + \ldots + (620.60)^2}{5} - \frac{(3176.00)^2}{25}$$

$$= 1850.83$$

Finally obtain the pooled unadjusted block sum of squares, SSB (unadj.), as

$$SSB\,(\text{unadj.}) = SSB_1(\text{unadj.}) + SSB_2\,(\text{unadj.}) \qquad (4.59)$$
$$= 1219.75 + 1850.83 = 3070.58$$

Compute the following correction quantity Q to be subtracted from the unadjusted treatment sum of squares :

$$Q = k(r-1)\mu\left[\left\{\frac{r}{(r-1)(1+k\mu)}\right\}(SSB\,(\text{unadj.}) - SSB\,(\text{adj.}))\right] \qquad (4.60)$$

where $\mu = \dfrac{E_b - E_e}{k(r-1)E_b}$ (4.61)

where E_b = Adjusted inter block mean square

E_e = Intra-block mean square

For our example, $\mu = \dfrac{227.44 - 442.36}{5(2-1)227.44}$

$= - 0.189$

$$Q = (5)(2-1)(-0.189)\left[\left\{\frac{2}{(2-1)(1+\{5\}\{-0.189\})}\right\}\{(3070.58)-(1819.56)\}\right]$$

$= -42989.60$

Finally, subtract this quantity Q from the unadjusted treatment sum of squares to obtain the adjusted sum of squares for treatment.

SST (adj.) = SST (unadj.) - Q (4.62)

$= 9087.29 - (-42989.60) = 52076.89$

Set up the following ANOVA table for testing the significance of the treatment effects.

Table 4.40. ANOVA table for testing the significance of adjusted treatment means.

Source of variation	Degree of freedom (df)	Sum of squares (SS)	Mean square $\left(MS = \frac{SS}{df}\right)$	Computed F	Tabular F
Treatment (adj.)	25	52076.89	2083.08	4.709	2.24
Intra-block error	16	7077.73	442.358		

The F-value computed for the present example has turned out to be significant at 5% level indicating that significant differences among the treatments. The sensitivity of F-test increased after eliminating the block effects. Generally, the block effect, as judged from E_b value is supposed to be greater than the intra-block error E_e. although this did not happen with the present example.

Adjustments are necessary to treatment means as well, since the ordinary treatment means are not unbiased estimates of true values of treatment means. The procedure for effecting such adjustments to eliminate block effects is as follows:

Step 10.Compute a correction term for each block by multiplying each C_b value by the quantity μ (-0.189), given by (4.61).

For Replication I, these values are :

μC_1 = -8.13, μC_2= -4.74, μC_3 = 6.01, μC_4 = -14.78, μC_5 = -20.22

For Replication II, these values are :

μC_6 =18.99, μC_7 = 14.59, μC_8 = 3.74, μC_9 =1.59, μC_{10} = 2.95

Enter these values in the last row and last column of Table 4.34 as shown in Table 4.31. Check that the total of all μC_b values add up to zero except for rounding of error. *i.e.*,

$\mu C_1 + \mu C_2 + \ldots + \mu C_{10}$= -8.13+-4.74 +…+ 2.95= 0.00

Make entries of the μC_b values for Replication I along the last column of Table 4.41 and μC_b values for Replication II along the last row of the same Table 4.41. This way of writing the correction values to be effected to the unadjusted totals for treatments will save a lot of confusion in carrying out arithmetical calculations. Each treatment total in Table 4.41 is now to be adjusted for block effects by applying the block corrections appropriate to the blocks in which that treatment appears.

Table 4. 41. Treatment totals and correction factors.

1 260.80	2 264.60	3 266.30	4 275.90	5 284.00	μC_1 = -8.13
6 284.70	7 244.70	8 246.20	9 214.10	10 223.80	μC_2= -4.74
11 220.30	12 210.60	13 242.60	14 196.90	15 229.00	μC_3 = 6.01
16 276.30	17 213.90	18 230.30	19 209.50	20 254.80	μC_4= -14.78
21 253.80	22 297.60	23 267.40	24 228.40	25 234.00	μC_5 = -20.22
μC_6 =18.99	μC_7 = 14.59	μC_8 = 3.74	μC_9 =1.59	μC_{10} = 2.95	

For example, clone 1 appears in Block 1 of Replication I and Block 6 of Replication 2. Add the values μC_1 and μC_6 to the total for clone 1.

i.e., The adjusted treatment total for clone 1 = 260.80 -(-8.13) - 18.99 = 2.55

Since the block corrections are already entered along the row and column in Table 4.41, the adjusted treatment totals are merely obtained by the respective column and row values for μC_b in which that treatment appears. Finally,

construct a table showing the treatment total adjusted for block effects. The adjusted values are shown in table 4.42 below.

Table 4.42. Adjusted treatment totals:

1 249.94	2 258.14	3 270.69	4 282.44	5 289.18
6 270.45	7 234.85	8 247.2	9 217.25	10 225.59
11 195.30	12 190.00	13 232.85	14 189.30	15 220.04
16 272.09	17 214.09	18 241.34	19 222.69	20 266.63
21 255.03	22 303.23	23 283.88	24 247.03	25 251.27

Obtain the adjusted treatment means by dividing each value by 2 since each total contains two observations from 2 replications (Table 4.43)

Table 4.43. Adjusted treatment means

1 124.97	2 129.07	3 135.35	4 141.22	5 144.59
6 135.23	7 117.43	8 123.60	9 108.63	10 112.80
11 97.65	12 95.00	13 116.43	14 94.65	15 110.02
16 136.05	17 107.05	18 120.67	19 111.35	20 133.32
21 127.52	22 151.62	23 141.94	24 123.52	25 125.64

4.7.3. Comparison of means

It was already mentioned that in a partially balanced lattice design, treatments that occur in the same block are compared with greater precision (*i.e.*, smaller standard error) than the treatments that occur in different blocks.

The formula for standard error for comparing any two treatment means that occur together in the same block is given by,

$$SE(d)_1 = \sqrt{\frac{2E_e}{r}\left[1+(r-1)\mu\right]} \qquad (4.63)$$

where $\mu = \dfrac{E_b - E_e}{k(r-1)E_b}$

E_b = Inter block mean square
E_e = Intra-block mean square
r = Number of replications

For our example,

$$SE(d)_1 = \sqrt{\frac{2 \times 442.3579}{2}\left[1+(2-1)(-0.189)\right]} = 18.9408$$

Standard error for comparing treatment means that occur in different blocks is,

$$SE(d)_2 = \sqrt{\frac{2E_e}{r}\left(1+r\mu\right)} \qquad (4.64)$$

For our example,

$$SE(d)_2 = \sqrt{\frac{2 \times 442.3579}{2}\left[1+2 \times (-0.189\right]} = 16.5875$$

Note that $SE(d)_2 < SE(d)_1$ in this example, because of the peculiarities of the data. This is not the usual case

These standard errors when multiplied by the tabular t value for the intra-block error degrees of freedom at the specified level of significance, will provide LSD value with which the adjusted treatment means can be compared for significant differences.

4.8. Response surface designs

In experiments where one or more quantitative factors are tested at multiple levels, it is often convenient to summarise the data by fitting a suitable model depicting the factor-response relationship. The quantitative factors may be fertiliser, irrigation, stand density etc., and the experiment may be to find out how the levels of these factors affect the response, γ. The response γ may be represented as a suitable function of the levels $x_{1u}, x_{2u}, \ldots, x_{ku}$ of the k factors and β, the set of parameters. A typical model may be

$$\gamma_u = f(x_{1u}, x_{2u}, \ldots, x_{ku}; \beta) + e_u \qquad (4.65)$$

where $u = 1, \ldots, t$ represents the N observations with x_{iu} representing the level of the ith factor ($i = 1, 2, \ldots, k$) in the in the uth observation. The residual e_u measures the experimental error of the uth observation. The function f is called the response surface. A knowledge of f gives a complete summary of the results of the experiment and helps in obtaining the optimum dose combination. It also enables prediction of the response for values of the x_{iu} that were not tested in the experiment. The designs specifically suited for fitting of response surfaces are known as response surface

designs. The response surfaces are usually approximated by polynomials of suitable degree, most common of which are second degree polynomials. Hence, designs suitable for fitting a second degree polynomial are described here.

4.8.1. Second order rotatable design

Let there be k factors such that ith factor has s_i levels. In all, there will be $s_1 x\ s_2 x \ldots\ldots x\ s_k$ treatment combinations out of which t combinations are taken to fit a second degree function of the form.

$$y_u = \beta_0 + \sum_i^k \beta_i x_{iu} + \sum_i^k \beta_{ii} x_{iu}^{\ 2} + \sum_{i<j}^k \beta_{ij} x_{iu} x_{ju} + e_u \qquad (4.66)$$

where y_u is the response obtained from the uth combination of factors (u = 1, 2, …, t)
x_{iu} is the level of the ith factor in the uth observation
β_0 is a constant
β_i is the ith linear regression coefficient
β_{ii} is the ith quadratic regression coefficient
β_{ij} is the (i,j)th interaction coefficient
e_u is the random error component associated with the uth observation with zero mean and constant variance.

For example, a specific case of model (4.66) involving only two factors would be,

$$y_u = \beta_0 + \beta_1 x_{1u} + \beta_2 x_{2u} + \beta_{11} x_{1u}^2 + \beta_{22} x_{2u}^2 + \beta_{12} x_{1u} x_{2u} + e_u$$

A second order response surface design enables the fitting of a second order polynomial for the factor-response relationship effectively. While choosing the design points, certain constraints are imposed on the levels of factors such that the parameter estimation gets simplified and also the resulting design and the model fitted through the design have certain desirable properties. One such property is the rotatability of the design. Rotatable designs make the variance of the estimated response from any treatment combination, as a function of the sum of squares of the levels of the factors in that treatment combination. Expressed alternatively, an experimental design is said to be rotatable if the variance of the predicted response at some specific set of x values is a function only of the distance of the point defined by the x values from the design centre and is not a function of the direction. It has been shown that the following conditions are necessary for the n design points to constitute a second order rotatable design (SORD).

(i) $$\sum_u x_{iu} = \sum_u x_{iu} x_{ju} = \sum_u x_{iu} x_{ju}^2 = \sum_u x_{iu}^3 = 0,$$
$$\sum_u x_{iu} x_{ju}^3 = \sum_u x_{iu} x_{ju} x_{ku}^2 = \sum_u x_{iu} x_{ju} x_{ku} = \sum_u x_{iu} x_{ju} x_{ku} x_{lu} = 0. \qquad (4.67)$$

(ii) $$\sum_u x_{iu}^2 = t l_2 \qquad (4.68)$$

(iii) $\sum_u x_{iu}^4 = 3t l_4$ (4.69)

(iv) $\sum_u x_{iu}^2 x_{ju}^2 = t l_4$ for $i \neq j$ or $\sum_u x_{iu}^4 = 3\sum_u x_{iu}^2 x_{ju}^2$ for $i \neq j$ (4.70)

(v) $\frac{\lambda_4}{\lambda_2^2} > \frac{k}{(k+2)}$ (4.71)

4.8.2. Construction of SORD

One of the commonly used methods for construction of SORD is given below which results in a class of designs by name central composite designs. Let there be k factors. A central composite design consists of a 2^k factorial or fractional factorial (coded to the usual ± 1 notation) augmented by $2k$ axial points, (± α, 0, 0,…, 0), (0, ± α, 0,…, 0), (0,0,± α, 0,…,0), …, (0,0,0,…,± α) and n_c centre points (0,0,…, 0). In case a fractional factorial is chosen for the first set of 2^k points, with $k > 4$, it must be seen that the defining contrasts do not involve any interaction with less than five factors. Central composite design for $k = 3$ is shown below. The design consists of $2^3 = 8$ factorial points, (2)(3) = 6 axial points and 1 centre point constituting a total of 15 points.

x_1	x_2	x_3
-1	-1	-1
-1	-1	+1
-1	+1	-1
-1	+1	+1
+1	-1	-1
+1	-1	+1
+1	+1	-1
+1	+1	+1
+α	0	0
-α	0	0
0	+α	0
0	-α	0
0	0	+α
0	0	-α
0	0	0

A central composite design is made rotatable by the choice of α. The value of α depends on the number points in the factorial portion of the design. In fact, $\alpha = (n_f)^{1/4}$ yields a rotatable central composite design where n_f is the number points used in the factorial portion of the design. In our example, the factorial portion contains $n_f = 2^3 = 8$ points. Thus the value of α for rotatability is = $(8)^{1/4}$ = 1.682. Additional details and examples of SORD can be found in Das and Giri (1979) and Montgomery (1991).

The treatment combinations specified by SORD may be tried with sufficient number of replications under any standard design in an experiment following the regular

randomization procedure. Response surface design thus pertains only to a particular way of selecting the treatment combination in a factorial experiment and not to any physical design used for experimental layout.

4.8.3. Fitting of a second degree response surface from a SORD

The analysis of data from a SORD laid out under completely randomized design is illustrated in the following. Let there be t distinct design points in the experiment with n_g replications for the gth design point. Let y_{gu} be the response obtained from the uth replication of the gth design point. Let x_{igu} be the level of the ith factor in the uth replication of the gth design point ($i = 1,\ldots, k$; $g = 1,\ldots, t$; $u = 1,\ldots n_g$). Let the total number of observations be n and $(p+1)$ be the number of parameters in the second order model to be fitted.

For the illustration of the analysis, data from a pot culture experiment is utilized. In order to simplify the discussion, certain modifications were made both in the data and design structure and to that extent, the data set is hypothetical. Nevertheless, the purpose of illustration is well-served by the example. The experiment included three factors *viz*., the quantity of nitrogen (N), phosphorus (P) and potassium (P) applied in the form urea, super phosphate and muriate of potash respectively. The experimental units were pots planted with two-year old seedlings of cane (*Calamus hookerianus*), each pot carrying a single seedling. The range of each element N, P and K included in the experiment was 5 g to 20 g/pot. The treatment structure corresponded to the central composite design discussed in Section 4.8.1, the physical design used being CRD with two replications. Since α=1.682 was the largest coded level in the design, the other dose levels were derived by equating α to 20g. Thus the other dose levels are $(-\alpha)$ = 5g, (-1) = 8.041g , (0) =12.5g, , (+1) =16.959g, (α) = 20g. The data obtained on oven-dry weight of shoot at the end of 2 years of the experiment are reported in Table 4.44.

Table 4.44. The data on oven-dry weight of shoot at the end of 2 years of the experiment.

N (x_1)	P (x_2)	K (x_3)	Shoot weight (g) (y)	
			Seedling 1	Seedling 2
-1	-1	-1	8.60	7.50
-1	-1	1	9.00	8.00
-1	1	-1	9.20	8.10
-1	1	1	11.50	9.10
1	-1	-1	10.00	9.20
1	-1	1	11.20	10.20
1	1	-1	11.00	9.90
1	1	1	12.60	11.50
1.682	0	0	11.00	10.10
-1.682	0	0	8.00	6.80
0	1.682	0	11.20	10.10
0	-1.682	0	9.50	8.50

0	0	1.682	11.50	10.50
0	0	-1.682	10.00	8.80
0	0	0	11.00	10.00

The steps involved in the analysis are the following.

Step 1. Calculate the values of λ_2 and λ_4 using Equations (4.68) and (4.69).

$15\lambda_2 = 13.65825$

$\lambda_2 = 0.9106$

$3t\lambda_4 = 24.00789$

$\lambda_4 = 0.5335$

As per the notation in Equations (4.68) and (4.69), t was taken as the number of distinct points in the design.

Step 2. Construct an outline of analysis of variance table as follows.

Table 4.45. Schematic representation of ANOVA table for fitting SORD.

Source of variation	Degree of freedom	Sum of squares	Mean square	Computed F
Regression	p	SSR	MSR	$\frac{MSR}{MSE}$
Lack of fit	n - 1- $\sum_{g=1}^{t}(n_g - 1)$ -p	SSL	MSL	$\frac{MSL}{MSE}$
Pure error	$\sum_{g=1}^{t}(n_g - 1)$	SSE	MSE	
Total	n - 1	SSTO		

Step 3. Compute the correction factor (*C.F.*)

$$C.F. = \frac{\left(\sum_{g=1}^{t}\sum_{u=1}^{n_g} y_{gu}\right)^2}{n} \qquad (4.72)$$

$$= \frac{(8.60 + 7.50 + ... + 10.00)^2}{30}$$

$$= 2873.37$$

Step 4. Compute the total sum of squares as

$$SSTO = \sum_{g=1}^{t}\sum_{u=1}^{n_g} y_{gu}^2 - C.F. \tag{4.73}$$

$$= (8.60)^2 + (7.50)^2 + \ldots + (10.00)^2 - \frac{(293.60)^2}{30}$$

$$= 55.43$$

Step 5. Compute the estimates of regression coefficients

$$\hat{\beta}_0 = \frac{\lambda_4(k+2)\sum_{g=1}^{t}\sum_{u=1}^{n_g} y_{gu} - \lambda_2\sum_{i=1}^{k}\sum_{g=1}^{t}\sum_{u=1}^{n_g} y_{gu}x_{igu}^2}{n\left[\lambda_4(k+2) - k\lambda_2^2\right]} \tag{4.74}$$

$$= \frac{(0.5335)(3+2)(293.60) - 0.9106(797.98)}{30\left[0.5335(3+2) - 3(0.9106)^2\right]}$$

$$= 10.47$$

$$\beta_i = \frac{\sum_{g=1}^{t}\sum_{u=1}^{n_g} y_{gu}x_{igu}}{n\lambda_2} \tag{4.75}$$

$$\beta_1 = \frac{25.20}{(30)(0.9106)}$$

$$= 0.92$$

$$\beta_2 = \frac{14.75}{(30)(0.9106)}$$

$$= 0.54$$

$$\beta_3 = \frac{14.98}{(30)(0.9106)}$$

$$= 0.55$$

$$\hat{\beta}_{ii} = \frac{1}{2n\lambda_4}\left(\sum_{g=1}^{t}\sum_{u=1}^{n_g} y_{gu}x_{igu}^2 + \frac{\left[\lambda_2^2 - \lambda_4\right]\sum_{i=1}^{k}\sum_{g=1}^{t}\sum_{u=1}^{n_g} y_{gu}x_{igu}^2 - 2\lambda_2\lambda_4\sum_{g=1}^{t}\sum_{u=1}^{n_g} y_{gu}}{\left[\lambda_4(k+2) - k\lambda_2^2\right]}\right) \tag{4.76}$$

$$\hat{\beta}_{11} = \frac{1}{(2)(30)(0.5335)}\left(258.17 + \frac{\left[(0.9106)^2 - 0.5335\right](797.98) - (2)(0.9106)(0.5335)(293.60)}{\left[(0.5335)(3+2) - (3)(0.9106)^2\right]}\right)$$

$$= -0.50$$

$$\hat{\beta}_{22} = \frac{1}{(2)(30)(0.5335)}\left(267.78 + \frac{\left[(0.9106)^2 - 0.5335\right](797.98) - (2)(0.9106)(0.5335)(293.60)}{\left[(0.5335)(3+2) - (3)(0.9106)^2\right]}\right)$$

$$= -0.20$$

$$\hat{\beta}_{33} = \frac{1}{(2)(30)(0.5335)}\left(272.03 + \frac{\left[(0.9106)^2 - 0.5335\right](797.98) - (2)(0.9106)(0.5335)(293.60)}{\left[(0.5335)(3+2) - (3)(0.9106)^2\right]}\right)$$

$$= -0.06$$

$$\hat{\beta}_{ij} = \frac{\sum_{g=1}^{t}\sum_{u=1}^{n_g} y_{gu} x_{igu} x_{jgu}}{n\lambda_4} \qquad (4.77)$$

$$\hat{\beta}_{12} = \frac{(-0.40)}{(30)(0.5335)}$$

$$= -0.02$$

$$\hat{\beta}_{13} = \frac{(1..20)}{(30)(0.5335)}$$

$$= 0.07$$

$$\hat{\beta}_{23} = \frac{(3.40)}{(30)(0.5335)}$$

$$= 0.21$$

Step 6. Compute the SSR as

$$SSR = \hat{\beta}_0 \sum\sum y_{gu} + \sum_i \hat{\beta}_i \sum_u y_{gu} x_{igu} + \sum_i \hat{\beta}_{ii} \sum_u y_{gu} x_{igu}^{\ 2} + \sum_{i<j} \hat{\beta}_{ij} \sum_u y_{gu} x_{igu} x_{jgu} - C.F \qquad (4.78)$$

$$= (10.47)(293.60) + (0.92)(25.20) + (0.54)(14.75) + (0.55)(14.98) + (-0.50)(258.17) +$$

$$(-0.20)(267.78) + (-0.06)(272.03) + (-0.02)(-0.40) + (0.07)(1.20) + (0.21)(3.40) -$$

$$\frac{(293.60)^2}{30}$$

$$= 44.42$$

Step 7. Calculate the sum of squares due to pure error

$$SSE = \sum_{g=1}^{t}\sum_{u=1}^{n_g}\left(y_{gu} - \bar{y}_g\right)^2 \qquad (4.79)$$

$$= 9.9650$$

Step 8. Calculate the lack of fit sum of squares as

$$SSL = SSTO - SSR - SSE \tag{4.80}$$

$$= 55.4347 - 44.4232 - 9.650$$
$$= 1.0465$$

Step 9. Enter the different sums of squares in the ANOVA table and compute the different mean squares by dividing the sums of squares by the respective degrees of freedom.

Table 4.46. ANOVA table for fitting SORD using data in Table 4.44

Source of variation	Degree of freedom	Sum of squares	Mean square	Computed *F*	Tabular *F* 5%
Regression	9	44.4232	4.9359	7.4299	2.56
Lack of fit	5	1.0465	0.2093	0.3150	2.90
Pure error	15	9.9650	0.6643		
Total	29	55.4347			

Step 10.Calculate the *F* value for testing significance of lack of fit which tests for the presence of any mis-specification in the model.

$$F = \frac{\text{Lack of fit } MS}{\text{Pure error } MS} \tag{4.81}$$

If the lack of fit is found significant, then the regression mean square is tested against the lack of fit mean square. Otherwise the regression mean square can be tested against the pure error mean square.

For our example $F = \dfrac{0.2093}{0.6643} = 0.3150$

Here, the lack of fit is found to be nonsignificant. Hence, the regression mean square can be tested against the pure error mean square. The *F* value for testing significance of regression is

$$F = \frac{\text{Regression } MS}{\text{Pure error } MS} \tag{4.82}$$
$$= \frac{4.9359}{0.6643}$$
$$= 7.4299$$

The *F* value for regression is significant when compared with tabular *F* value of 2.56 for 9 and 15 degrees of freedom at 5 % level of significance. The model was found to explain nearly 80 % of the variation in the response variable as

could be seen from the ratio of the regression sum of squares to the total sum of squares.

Step 11. The variances and covariances of the estimated coefficients are obtained by

$$V\left(\hat{\beta}_0\right) = \frac{\lambda_4(k+2)}{n\left[\lambda_4(k+2) - k\lambda_2^2\right]}E \tag{4.83}$$

$$= \frac{(0.5335)(3+2)}{30\left[(0.5335)(3+2) - 3(0.9106)^2\right]}(0.6643)$$

$$= 0.3283$$

where E = Pure error mean square in the ANOVA table.

$$V(\hat{\beta}_i) = \frac{E}{n\lambda_2} \tag{4.84}$$

$$= \frac{0.6643}{(30)(0.9106)}$$

$$= 0.0243$$

$$V(\hat{\beta}_{ii}) = \frac{E}{2n\lambda_2}\left(1 + \frac{\left[\lambda_2^2 - \lambda_4\right]}{\left[\lambda_4(k+2) - k\lambda_2^2\right]}\right) \tag{4.85}$$

$$= \frac{0.6643}{(2)(30)(0.9106)}\left(1 + \frac{\left[(0.9106)^2 - 0.5335\right]}{\left[(0.5335)(3+2) - (3)(0.9106)^2\right]}\right)$$

$$= 0.03$$

$$V(\hat{\beta}_{ij}) = \frac{E}{n\lambda_4} \tag{4.86}$$

$$= \frac{0.6643}{(30)(0.5335)}$$

$$Cov(\hat{\beta}_0, \hat{\beta}_{ii}) = \frac{-\lambda_2}{n\left[\lambda_4(k+2) - k\lambda_2^2\right]}E \tag{4.87}$$

$$= \frac{-0.5335}{30\left[(0.5335)(3+2) - (3)(0.9106)^2\right]}(0.6643)$$

$$= -0.11$$

$$Cov(\hat{\beta}_{ii}, \hat{\beta}_{jj}) = \frac{\left[\lambda_2^2 - \lambda_4\right]}{2n\lambda_4\left[\lambda_4(k+2) - k\lambda_2^2\right]} \tag{4.88}$$

$$= \frac{\left((0.9106)^2 - 0.5335\right)}{(2)(30)(0.5335)\left[(0.5335)(3+2) - (3)(0.9106)^2\right]}$$

$$= 0.05$$

All other covariances will be zero.

The fitted response function, therefore is

$$\hat{y} = \hat{\beta}_0 + \sum_i \hat{\beta}_i x_i + \sum_i \hat{\beta}_{ii} x_i^2 + \sum_{i<j} \hat{\beta}_{ij} x_i x_j$$
$$= 10.47 + 0.92x_1 + 0.54x_2 + 0.55x_3 - 0.50x_1^2 - 0.20x_2^2 - 0.02x_3^2$$
$$- 0.02x_1x_2 + 0.07x_1x_3 + 0.21x_2x_3$$

One of the uses of the surface is to obtain the optimum dose combination at which the response is either maximum or economically optimum. Also with the help of the fitted equation, it is possible to investigate the nature of the surface in specific ranges of the input variables. Since the treatment of these aspects requires knowledge of advanced mathematical techniques, further discussion is avoided here but the details can be found in Montgomery (1991).

5. SAMPLING TECHNIQUES

5.1. Basic concepts of sampling

Essentially, sampling consists of obtaining information from only a part of a large group or population so as to infer about the whole population. The object of sampling is thus to secure a sample which will represent the population and reproduce the important characteristics of the population under study as closely as possible.

The principal advantages of sampling as compared to complete enumeration of the population are reduced cost, greater speed, greater scope and improved accuracy. Many who insist that the only accurate way to survey a population is to make a complete enumeration, overlook the fact that there are many sources of errors in a complete enumeration and that a hundred per cent enumeration can be highly erroneous as well as nearly impossible to achieve. In fact, a sample can yield more accurate results because the sources of errors connected with reliability and training of field workers, clarity of instruction, mistakes in measurement and recording, badly kept measuring instruments, misidentification of sampling units, biases of the enumerators and mistakes in the processing and analysis of the data can be controlled more effectively. The smaller size of the sample makes the supervision more effective. Moreover, it is important to note that the precision of the estimates obtained from certain types of samples can be estimated from the sample itself. The net effect of a sample survey as compared to a complete enumeration is often a more accurate answer achieved with fewer personnel and less work at a low cost in a short time.

The most ‘convenient’ method of sampling is that in which the investigator selects a number of sampling units which he considers ‘representative’ of the whole population.

For example, in estimating the whole volume of a forest stand, he may select a few trees which may appear to be of average dimensions and typical of the area and measure their volume. A walk over the forest area with an occasional stop and flinging a stone with the eyes closed or some other simple way that apparently avoids any deliberate choice of the sampling units is very tempting in its simplicity. However, it is clear that such methods of selection are likely to be biased by the investigator's judgement and the results will thus be biased and unreliable. Even if the investigator can be trusted to be completely objective, considerable conscious or unconscious errors of judgement, not frequently recognized, may occur and such errors due to bias may far outweigh any supposed increase in accuracy resulting from deliberate or purposive selection of the units. Apart from the above points, subjective sampling does not permit the evaluation of the precision of the estimates calculated from samples. Subjective sampling is statistically unsound and should be discouraged.

When sampling is performed so that every unit in the population has some chance of being selected in the sample and the probability of selection of every unit is known, the method of sampling is called probability sampling. An example of probability sampling is random selection, which should be clearly distinguished from haphazard selection, which implies a strict process of selection equivalent to that of drawing lots. In this manual, any reference to sampling, unless otherwise stated, will relate to some form of probability sampling. The probability that any sampling unit will be selected in the sample depends on the sampling procedure used. The important point to note is that the precision and reliability of the estimates obtained from a sample can be evaluated only for a probability sample. Thus the errors of sampling can be controlled satisfactorily in this case.

The object of designing a sample survey is to minimise the error in the final estimates. Any forest survey involving data collection and analysis of the data is subject to a variety of errors. The errors may be classified into two groups *viz*., (i) non-sampling errors (ii) sampling errors. The non-sampling errors like the errors in location of the units, measurement of the characteristics, recording mistakes, biases of enumerators and faulty methods of analysis may contribute substantially to the total error of the final results to both complete enumeration and sample surveys. The magnitude is likely to be larger in complete enumeration since the smaller size of the sample project makes it possible to be more selective in assignment of personnel for the survey operations, to be more thorough in their training and to be able to concentrate to a much greater degree on reduction of non-sampling errors. Sampling errors arise from the fact that only a fraction of the forest area is enumerated. Even if the sample is a probability sample, the sample being based on observations on a part of the population cannot, in general, exactly represent the population. The average magnitude of the sampling errors of most of the probability samples can be estimated from the data collected. The magnitude of the sampling errors, depends on the size of the sample, the variability within the population and the sampling method adopted. Thus if a probability sample is used, it is possible to predetermine the size of the sample needed to obtain desired and specified degree of precision.

A sampling scheme is determined by the size of sampling units, number of sampling units to be used, the distribution of the sampling units over the entire area to be sampled, the type and method of measurement in the selected units and the statistical procedures for analysing the survey data. A variety of sampling methods and estimating techniques developed to meet the varying demands of the survey statistician accord the user a wide selection for specific situations. One can choose the method or combination of methods that will yield a desired degree of precision at minimum cost. Additional references are Chacko (1965) and Sukhatme *et al, (1984)*

5.1.1. The principal steps in a sample survey

In any sample survey, we must first decide on the type of data to be collected and determine how adequate the results should be. Secondly, we must formulate the sampling plan for each of the characters for which data are to be collected. We must also know how to combine the sampling procedures for the various characters so that no duplication of field work occurs. Thirdly, the field work must be efficiently organised with adequate provision for supervising the work of the field staff. Lastly, the analysis of the data collected should be carried out using appropriate statistical techniques and the report should be drafted giving full details of the basic assumptions made, the sampling plan and the results of the statistical analysis. The report should contain estimate of the margin of the sampling errors of the results and may also include the possible effects of the non-sampling errors. Some of these steps are elaborated further in the following.

(i) Specification of the objectives of the survey: Careful consideration must be given at the outset to the purposes for which the survey is to be undertaken. For example, in a forest survey, the area to be covered should be decided. The characteristics on which information is to be collected and the degree of detail to be attempted should be fixed. If it is a survey of trees, it must be decided as to what species of trees are to be enumerated, whether only estimation of the number of trees under specified diameter classes or, in addition, whether the volume of trees is also proposed to be estimated. It must also be decided at the outset what accuracy is desired for the estimates.

(*ii*) *Construction of a frame of units :* The first requirement of probability sample of any nature is the establishment of a frame. The structure of a sample survey is determined to a large extent by the frame. A frame is a list of sampling units which may be unambiguously defined and identified in the population. The sampling units may be compartments, topographical sections, strips of a fixed width or plots of a definite shape and size.

The construction of a frame suitable for the purposes of a survey requires experience and may very well constitute a major part of the work of planning the survey. This is particularly true in forest surveys since an artificial frame composed of sampling units of topographical sections, strips or plots may have to be constructed. For instance, the basic component of a sampling frame in a forest survey may be a proper map of the forest area. The choice of sampling units must be one that permits the identification in the field of a particular sampling unit which has to be selected in the sample. In forest

surveys, there is considerable choice in the type and size of sampling units. The proper choice of the sampling units depends on a number of factors; the purpose of the survey, the characteristics to be observed in the selected units, the variability among sampling units of a given size, the sampling design, the field work plan and the total cost of the survey. The choice is also determined by practical convenience. For example, in hilly areas it may not be practicable to take strips as sampling units. Compartments or topographical sections may be more convenient. In general, at a given intensity of sampling (proportion of area enumerated) the smaller the sampling units employed the more representative will be the sample and the results are likely to be more accurate.

(iii) Choice of a sampling design: If it is agreed that the sampling design should be such that it should provide a statistically meaningful measure of the precision of the final estimates, then the sample should be a probability sample, in that every unit in the population should have a known probability of being selected in the sample. The choice of units to be enumerated from the frame of units should be based on some objective rule which leaves nothing to the opinion of the field worker. The determination of the number of units to be included in the sample and the method of selection is also governed by the allowable cost of the survey and the accuracy in the final estimates.

(iv) Organisation of the field work : The entire success of a sampling survey depends on the reliability of the field work. In forest surveys, the organization of the field work should receive the utmost attention, because even with the best sampling design, without proper organization the sample results may be incomplete and misleading. Proper selection of the personnel, intensive training, clear instructions and proper supervision of the fieldwork are essential to obtain satisfactory results. The field parties should correctly locate the selected units and record the necessary measurements according to the specific instruction given. The supervising staff should check a part of their work in the field and satisfy that the survey carried out in its entirety as planned.

(v) Analysis of the data : Depending on the sampling design used and the information collected, proper formulae should be used in obtaining the estimates and the precision of the estimates should be computed. Double check of the computations is desired to safeguard accuracy in the analysis.

(vi) Preliminary survey (pilot trials) : The design of a sampling scheme for a forest survey requires both knowledge of the statistical theory and experience with data regarding the nature of the forest area, the pattern of variability and operational cost. If prior knowledge in these matters is not available, a statistically planned small scale 'pilot survey' may have to be conducted before undertaking any large scale survey in the forest area. Such exploratory or pilot surveys will provide adequate knowledge regarding the variability of the material and will afford opportunities to test and improve field procedures, train field workers and study the operational efficiency of a design. A pilot survey will also provide data for estimating the various components of cost of operations in a survey like time of travel, time of location and enumeration of sampling units, etc. The above information will be of great help in deciding the proper type of design and intensity of sampling that will be appropriate for achieving the objects of the survey.

5.1.2. Sampling terminology

Although the basic concepts and steps involved in sampling are explained above, some of the general terms involved are further clarified in this section so as to facilitate the discussion on individual sampling schemes dealt with in later sections.

Population : The word population is defined as the aggregate of units from which a sample is chosen. If a forest area is divided into a number of compartments and the compartments are the units of sampling, these compartments will form the population of sampling units. On the other hand, if the forest area is divided into, say, a thousand strips each 20 m wide, then the thousand strips will form the population. Likewise if the forest area is divided into plots of, say, one-half hectare each, the totality of such plots is called the population of plots.

Sampling units : Sampling units may be administrative units or natural units like topographical sections and subcompartments or it may be artificial units like strips of a certain width, or plots of a definite shape and size. The unit must be a well defined element or group of elements identifiable in the forest area on which observations on the characteristics under study could be made. The population is thus sub-divided into suitable units for the purpose of sampling and these are called sampling units.
Sampling frame : A list of sampling units will be called a 'frame'. A population of units is said to be finite if the number units in it is finite.

Sample : One or more sampling units selected from a population according to some specified procedure will constitute a sample.

Sampling intensity : Intensity of sampling is defined as the ratio of the number of units in the sample to the number of units in the population.

Population total : Suppose a finite population consists of units U_1, U_2, …, U_N. Let the value of the characteristic for the ith unit be denoted by y_i. For example the units may be strips and the characteristic may be the number of trees of a certain species in a strip. The total of the values y_i ($i = 1, 2, \ldots, N$), namely,

$$Y = \sum_{i=1}^{N} y_i \qquad (5.1)$$

is called the population total which in the above example is the total number of trees of the particular species in the population.

Population mean : The arithmetic mean

$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i \qquad (5.2)$$

is called the population mean which, in the example considered, is the average number of trees of the species per strip.

Population variance : A measure of the variation between units of the population is provided by the population variance

$$S_y^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{Y})^2 = \frac{1}{N}\sum_{i=1}^{N} y_i^2 - \bar{Y}^2 \qquad (5.3)$$

which in the example considered measures the variation in number of trees of the particular species among the strips. Large values of the population variance indicate large variation between units in the population and small values indicate that the values of the characteristic for the units are close to the population mean. The square root of the variance is known as *standard deviation.*

Coefficient of variation : The ratio of the standard deviation to the value of the mean is called the coefficient of variation, which is usually expressed in percentage.

$$C.\,V. = \frac{S_y}{\bar{Y}} \qquad (5.4)$$

The coefficient of variation, being dimensionless, is a valuable tool to compare the variation between two or more populations or sets of observations.

Parameter : A function of the values of the units in the population will be called a parameter. The population mean, variance, coefficient of variation, etc., are examples of population parameters. The problem in sampling theory is to estimate the parameters from a sample by a procedure that makes it possible to measure the precision of the estimates.

Estimator, estimate : Let us denote the sample observations of size n by $y_1, y_2, \ldots, y_n$. Any function of the sample observations will be called a *statistic*. When a statistic is used to estimate a population parameter, the statistic will be called an estimator. For example, the sample mean is an estimator of the population mean. Any particular value of an estimator computed from an observed sample will be called an estimate.

Bias in estimation : A statistic t is said to be an unbiased estimator of a population parameter q if its expected value, denoted by E(t), is equal to q. A sampling procedure based on a probability scheme gives rise to a number of possible samples by repetition of the sampling procedure. If the values of the statistic t are computed for each of the possible samples and if the average of the values is equal to the population value q, then t is said to be an unbiased estimator of q based on sampling procedure. Notice that the repetition of the procedure and computing the values of t for each sample is only conceptual, not actual, but the idea of generating all possible estimates by repetition of the sampling process is fundamental to the study of bias and of the assessment of sampling error. In case E(t) is not equal to q, the statistic t is said to be a biased estimator of q and the bias is given by, bias = E(t) - q. The introduction of a genuinely random process in selecting a sample is an important step in avoiding bias. Samples selected subjectively will usually be very seriously biased. In forest surveys, the tendency of forest officers to select typical forest areas for enumerations, however honest the intention may be, is bound to result in biased estimates.

Sampling variance : The difference between a sample estimate and the population value is called the sampling error of the estimate, but this is naturally unknown since the population value is unknown. Since the sampling scheme gives rise to different possible samples, the estimates will differ from sample to sample. Based on these possible estimates, a measure of the average magnitude over all possible samples of the squares of the sampling error can be obtained and is known as the *mean square error* (*MSE*) of the estimate which is essentially a measure of the divergence of an estimator from the true population value. Symbolically, $MSE = E[t - q]^2$. The sampling variance ($V(t)$)is a measure of the divergence of the estimate from its expected value. It is defined as the average magnitude over all possible samples of the squares of deviations of the estimator from its expected value and is given by $V(t) = E[t - E(t)]^2$.

Notice that the sampling variance coincides with the mean square error when t is an unbiased estimator. Generally, the magnitude of the estimate of the sampling variance computed from a sample is taken as indicating whether a sample estimate is useful for the purpose. The larger the sample and the smaller the variability between units in the population, the smaller will be the sampling error and the greater will be the confidence in the results.

Standard error of an estimator : The square root of the sampling variance of an estimator is known as the standard error of the estimator. The standard error of an estimate divided by the value of the estimate is called relative standard error which is usually expressed in percentage.

Accuracy and precision : The standard error of an estimate, as obtained from a sample, does not include the contribution of the bias. Thus we may speak of the standard error or the sampling variance of the estimate as measuring on the inverse scale, the precision of the estimate, rather than its *accuracy*. Accuracy usually refers to the size of the deviations of the sample estimate from the mean $m = E(t)$ obtained by repeated application of the sampling procedure, the bias being thus measured by $m - q$.

It is the accuracy of the sample estimate in which we are chiefly interested; it is the precision with which we are able to measure in most instances. We strive to design the survey and attempt to analyse the data using appropriate statistical methods in such a way that the precision is increased to the maximum and bias is reduced to the minimum.

Confidence limits : If the estimator t is normally distributed (which assumption is generally valid for large samples), a confidence interval defined by a lower and upper limit can expected to include the population parameter q with a specified probability level. The limits are given by

$$\text{Lower limit} = t - z\sqrt{\hat{V}(t)} \tag{5.5}$$

$$\text{Upper limit} = t + z\sqrt{\hat{V}(t)} \tag{5.6}$$

where $\hat{V}(t)$ is the estimate of the variance of t and z is the value of the normal deviate corresponding to a desired P % confidence probability. For example, when z is taken as 1.96, we say that the chance of the true value of q being contained in the random interval defined by the lower and upper confidence limits is 95 per cent. The confidence limits specify the range of variation expected in the population mean and also stipulate the degree of confidence we should place in our sample results. If the sample size is less than 30, the value of k in the formula for the lower and upper confidence limits should be taken from the percentage points of Student's t distribution (See Appendix 2) with degrees of freedom of the sum of squares in the estimate of the variance of t. Moderate departures of the distribution from normality does not affect appreciably the formula for the confidence limits. On the other hand, when the distribution is very much different from normal, special methods are needed. For example, if we use small area sampling units to estimate the average number of trees in higher diameter classes, the distribution may have a large skewness. In such cases, the above formula for calculating the lower and upper confidence limits may not be directly applicable.

Some general remarks : In the sections to follow, capital letters will usually be used to denote population values and small letters to denote sample values. The symbol 'cap' (^) above a symbol for a population value denotes its estimate based on sample observations. Other special notations used will be explained as and when they are introduced.

While describing the different sampling methods below, the formulae for estimating only population mean and its sampling variance are given. Two related parameters are population total and ratio of the character under study (y) to some auxiliary variable (x). These related statistics can always be obtained from the mean by using the following general relations.

$$\hat{Y} = N\hat{\bar{Y}} \tag{5.7}$$

$$V(\hat{Y}) = N^2 V(\hat{\bar{Y}}) \tag{5.8}$$

$$\hat{R} = \frac{\hat{Y}}{X} \tag{5.9}$$

$$V(\hat{R}) = \frac{V(\hat{Y})}{X^2} \tag{5.10}$$

where $\hat{Y}$ = Estimate of the population total
N = Total number of units in the population
$\hat{R}$ = Estimate of the population ratio
X = Population total of the auxiliary variable

5.2. Simple random sampling

A sampling procedure such that each possible combination of sampling units out of the population has the same chance of being selected is referred to as simple random

sampling. From theoretical considerations, simple random sampling is the simplest form of sampling and is the basis for many other sampling methods. Simple random sampling is most applicable for the initial survey in an investigation and for studies which involve sampling from a small area where the sample size is relatively small. When the investigator has some knowledge regarding the population sampled, other methods which are likely to be more efficient and convenient for organising the survey in the field, may be adopted. The irregular distribution of the sampling units in the forest area in simple random sampling may be of great disadvantage in forest areas where accessibility is poor and the costs of travel and locating the plots are considerably higher than the cost of enumerating the plot.

5.2.1. Selection of sampling units

In practice, a random sample is selected unit by unit. Two methods of random selection for simple random sampling without replacement are explained in this section.

(i) Lottery method : The units in the population are numbered 1 to *N*. If *N* identical counters with numberings 1 to *N* are obtained and one counter is chosen at random after shuffling the counters, then the probability of selecting any counter is the same for all the counters. The process is repeated *n* times without replacing the counters selected. The units which correspond to the numbers on the chosen counters form a simple random sample of size *n* from the population of *N* units.

(ii) Selection based on random number tables : The procedure of selection using the lottery method, obviously becomes rather inconvenient when *N* is large. To overcome this difficulty, we may use a table of random numbers such as those published by Fisher and Yates (1963) a sample of which is given in Appendix 6. The tables of random numbers have been developed in such a way that the digits 0 to 9 appear independent of each other and approximately equal number of times in the table. The simplest way of selecting a random sample of required size consists in selecting a set of *n* random numbers one by one, from 1 to *N* in the random number table and, then, taking the units bearing those numbers. This procedure may involve a number of rejections since all the numbers more than *N* appearing in the table are not considered for selection. In such cases, the procedure is modified as follows. If *N* is a *d* digited number, we first determine the highest *d* digited multiple of *N*, say *N'*. Then a random number *r* is chosen from 1 to *N'* and the unit having the serial number equal to the remainder obtained on dividing *r* by *N*, is considered as selected. If remainder is zero, the last unit is selected. A numerical example is given below.

Suppose that we are to select a simple random sample of 5 units from a serially numbered list of 40 units. Consulting Appendix 6 : Table of random numbers, and taking column (5) containing two-digited numbers, the following numbers are obtained:

39, 27, 00, 74, 07

In order to give equal chances of selection to all the 100 units, we are to reject all numbers above 79 and consider (00) equivalent to 80. We now divide the above numbers in turn by 40 and take the remainders as the selected strip numbers for our sample, rejecting the remainders that are repeated. We thus get the following 16 strip numbers as our sample :

39, 27, 40, 34, 7.

5.2.2. Parameter estimation

Let $y_1, y_2, \ldots, y_n$ be the measurements on a particular characteristic on n selected units in a sample from a population of N sampling units. It can be shown in the case of simple random sampling without replacement that the sample mean,

$$\hat{\bar{Y}} = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{5.11}$$

is an unbiased estimator of the population mean, $\bar{Y}$. An unbiased estimate of the sampling variance of $\bar{y}$ is given by

$$\hat{V}(\hat{\bar{Y}}) = \frac{N-n}{Nn} s_y^2 \tag{5.12}$$

where $s_y^2 = \dfrac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$ (5.13)

Assuming that the estimate $\bar{y}$ is normally distributed, a confidence interval on the population mean $\bar{Y}$ can be set with the lower and upper confidence limits defined by,

$$\text{Lower limit } \hat{\bar{Y}}_L = \bar{y} - z\frac{s_y}{\sqrt{n}}\sqrt{\frac{N-n}{N}} \tag{5.14}$$

$$\text{Upper limit } \hat{\bar{Y}}_U = \bar{y} + z\frac{s_y}{\sqrt{n}}\sqrt{\frac{N-n}{N}} \tag{5.15}$$

where z is the table value which depends on how many observations there are in the sample. If there are 30 or more observations we can read the values from the table of the normal distribution (Appendix 1). If there are less than 30 observations, the table value should be read from the table of t distribution (Appendix 2), using n - 1 degree of freedom.

The computations are illustrated with the following example. Suppose that a forest has been divided up into 1000 plots of 0.1 hectare each and a simple random sample of 25 plots has been selected. For each of these sample plots the wood volumes in m^3 were recorded. The wood volumes were,

7	10	7	4	7
8	8	8	7	5

2	6	9	7	8
6	7	11	8	8
7	3	8	7	7

If the wood volume on the *i*th sampling unit is designated as y_i , an unbiased estimator of the population mean, $\overline{Y}$ is obtained using Equation (5.11) as,

$$\hat{\overline{Y}} = \overline{y} = \frac{7+8+2+\ \ldots\ +7}{25} = \frac{175}{25}$$
$$= 7 \text{ m}^3$$

which is the mean wood volume per plot of 0.1 ha in the forest area.

An estimate (s_y^2) of the variance of individual values of y is obtained using Equation (5.13).

$$s_y^2 = \frac{(7-7)^2 + (8-7)^2 + \ \ldots + (7-7)^2}{25-1}$$
$$= \frac{82}{24} = 3.833$$

Then unbiased estimate of sampling variance of $\overline{y}$ is

$$\hat{V}(\hat{\overline{Y}}) = \left(\frac{1000-25}{(1000)(25)}\right)3.833$$
$$= 0.1495 \text{ (m}^3)^2$$

$$SE(\hat{\overline{Y}}) = \sqrt{0.1495} = 0.3867 \text{ m}^3$$

The relative standard error which is $\frac{SE(\hat{\overline{Y}})}{\hat{\overline{Y}}}(100)$ is a more common expression. Thus,

$$RSE(\hat{\overline{Y}}) = \frac{\sqrt{0.1495}}{7}(100) = 5.52\ \%$$

The confidence limits on the population mean $\overline{Y}$ are obtained using Equations (5.14) and (5.15).

$$\text{Lower limit } \hat{\overline{Y}}_L = 7 - (2.064)\sqrt{0.1495}$$
$$= 6.20 \text{ cords}$$
$$\text{Upper limit } \hat{\overline{Y}}_U = 7 + (2.064)\sqrt{0.1495}$$
$$= 7.80 \text{ cords}$$

The 95% confidence interval for the population mean is (6.20, 7.80) m^3. Thus, we are 95% confident that the confidence interval (6.20, 7.80) m^3 would include the population mean.

An estimate of the total wood volume in the forest area sampled can easily be obtained by multiplying the estimate of the mean by the total number of plots in the population. Thus,

$$\hat{Y} = 7(1000) = 7000 \text{ m}^3$$

with a confidence interval of (6200, 7800) obtained by multiplying the confidence limits on the mean by $N = 1000$. The RSE of $\hat{Y}$, however, will not be changed by this operation.

5.3. Systematic sampling

Systematic sampling employs a simple rule of selecting every kth unit starting with a number chosen at random from 1 to k as the random start. Let us assume that N sampling units in the population are numbered 1 to N. To select a systematic sample of n units, we take a unit at random from the first k units and then every kth sampling unit is selected to form the sample. The constant k is known as the *sampling interval* and is taken as the integer nearest to N / n, the inverse of the sampling fraction. Measurement of every kth tree along a certain compass bearing is an example of systematic sampling. A common sampling unit in forest surveys is a narrow strip at right angles to a base line and running completely across the forest. If the sampling units are strips, then the scheme is known as systematic sampling by strips. Another possibility is known as systematic line plot sampling where plots of a fixed size and shape are taken at equal intervals along equally spaced parallel lines. In the latter case, the sample could as well be systematic in two directions.

Systematic sampling certainly has an intuitive appeal, apart from being easier to select and carry out in the field, through spreading the sample evenly over the forest area and ensuring a certain amount of representation of different parts of the area. This type of sampling is often convenient in exercising control over field work. Apart from these operational considerations, the procedure of systematic sampling is observed to provide estimators more efficient than simple random sampling under normal forest conditions. The property of the systematic sample in spreading the sampling units evenly over the population can be taken advantage of by listing the units so that homogeneous units are put together or such that the values of the characteristic for the units are in ascending or descending order of magnitude. For example, knowing the fertility trend of the forest area the units (for example strips) may be listed along the fertility trend.

If the population exhibits a regular pattern of variation and if the sampling interval of the systematic sample coincides with this regularity, a systematic sample will not give precise estimates. It must, however, be mentioned that no clear case of periodicity has been reported in a forest area. But the fact that systematic sampling may give poor precision when unsuspected periodicity is present should not be lost sight of when planning a survey.

5.3.1. Selection of a systematic sample

To illustrate the selection of a systematic sample, consider a population of N = 48 units. A sample of n = 4 units is needed. Here, k = 12. If the random number selected from the set of numbers from 1 to 12 is 11, then the units associated with serial numbers 11, 23, 35 and 47 will be selected. In situations where N is not fully divisible by n, k is calculated as the integer nearest to N / n. In this situation, the sample size is not necessarily n and in some cases it may be n -1.

5.3.2. Parameter estimation

The estimate for the population mean per unit is given by the sample mean

$$\hat{\bar{Y}} = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \qquad (5.16)$$

where n is the number of units in the sample.

In the case of systematic strip surveys or, in general, any one dimensional systematic sampling, an approximation to the standard error may be obtained from the differences between pairs of successive units. If there are n units enumerated in the systematic sample, there will be (n-1) differences. The variance per unit is therefore, given by the sum of squares of the differences divided by twice the number of differences. Thus if $y_1, y_2,\ldots,y_n$ are the observed values (say volume) for the n units in the systematic sample and defining the first difference $d(y_i)$ as given below,

$$d(y_i) = y_{(i+1)} - y_{(i)}; \quad (i = 1, 2, \ldots, n-1), \qquad (5.17)$$

the approximate variance per unit is estimated as

$$\hat{V}(\hat{\bar{Y}}) = \frac{1}{2n(n-1)} \sum_{i=1}^{n-1} \left[d(y_i) \right]^2 \qquad (5.18)$$

As an example, Table 5.1 gives the observed diameters of 10 trees selected by systematic selection of 1 in 20 trees from a stand containing 195 trees in rows of 15 trees. The first tree was selected as the 8th tree from one of the outside edges of the stand starting from one corner and the remaining trees were selected systematically by taking every 20th tree switching to the nearest tree of the next row after the last tree in any row is encountered.

Table 5.1. Tree diameter recorded on a systematic sample of 10 trees from a plot.

Selected tree number	Diameter at breast-height(cm) y_i	First difference $d(y_i)$
8	14.8	
28	12.0	-2.8

48	13.6	+1.6
68	14.2	+0.6
88	11.8	-2.4
108	14.1	+2.3
128	11.6	-2.5
148	9.0	-2.6
168	10.1	+1.1
188	9.5	-0.6

Average diameter is equal to

$$\hat{\bar{Y}} = \frac{1}{10}(14.8 + 12.0 + \ldots\ldots + 9.5) = 12.07$$

The nine first differences can be obtained as shown in column (3) of the Table 5.1. The error variance of the mean per unit is thus

$$\hat{V}(\hat{\bar{Y}}) = \frac{(-2.8)^2 + (1.6)^2 + \ldots + (-0.6)^2}{2x9x10} = \frac{36.9}{180}$$
$$= 0.202167$$

A difficulty with systematic sampling is that one systematic sample by itself will not furnish valid assessment of the precision of the estimates. With a view to have valid estimates of the precision, one may resort to partially systematic samples. A theoretically valid method of using the idea of systematic samples and at the same time leading to unbiased estimates of the sampling error is to draw a minimum of two systematic samples with independent random starts. If $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_m$ are m estimates of the population mean based on m independent systematic samples, the combined estimate is

$$\bar{y} = \frac{1}{m}\sum_{i=1}^{m} \bar{y}_i \qquad (5.19)$$

The estimate of the variance of $\bar{y}$ is given by

$$\hat{V}(\bar{y}) = \frac{1}{m(m-1)}\sum_{i=1}^{m} (\bar{y}_i - \bar{y})^2 \qquad (5.20)$$

Notice that the precision increases with the number of independent systematic samples.

As an example, consider the data given in Table 5.1 along with another systematic sample selected with independent random starts. In the second sample, the first tree was selected as the 10th tree. Data for the two independent samples are given in Table 5.2.

Table 5.2. Tree diameter recorded on two independent systematic samples of 10 trees from a plot.

Sample 1	Sample 2

Selected tree number	Diameter at breast-height(cm) y_i	Selected tree number	Diameter at breast-height(cm) y_i
8	14.8	10	13.6
28	12.0	30	10.0
48	13.6	50	14.8
68	14.2	70	14.2
88	11.8	90	13.8
108	14.1	110	14.5
128	11.6	130	12.0
148	9.0	150	10.0
168	10.1	170	10.5
188	9.5	190	8.5

The average diameter for the first sample is $\bar{y}_1 = 12.07$. The average diameter for the first sample is $\bar{y}_2 = 12.19$. Combined estimate of population mean ($\bar{y}$) is obtained by using Equation (5.19) as,

$$\bar{y} = \frac{1}{2}(12.07 + 12.19)$$
$$= 12.13$$

The estimate of the variance of $\bar{y}$ is obtained by using Equation (5.20).

$$\hat{V}(\bar{y}) = \frac{1}{2(2-1)}(12.07 - 12.13)^2(12.19 - 12.13)^2 = 0.0036$$

$$SE(\bar{y}) = \sqrt{0.0036} = 0.06$$

One additional variant of systematic sampling is that sampling may as well be systematic in two directions. For example, in plantations, a systematic sample of rows and measurements on every tenth tree in each selected row may be adopted with a view to estimate the volume of the stand. In a forest survey, one may take a series of equidistant parallel strips extending over the whole width of the forest and the enumeration in each strip may be done by taking a systematic sample of plots or trees in each strip. Forming rectangular grids of (p x q) metres and selecting a systematic sample of rows and columns with a fixed size plot of prescribed shape at each intersection is another example.

In the case of two dimensional systematic sample, a method of obtaining the estimates and approximation to the sampling error is based on stratification and the method is similar to the stratified sampling given in section 5.4. For example, the sample may be arbitrarily divided into sets of four in 2 x 2 units and each set may be taken to form a stratum with the further assumption that the observations within each stratum are

independently and randomly chosen. With a view to make border adjustments, overlapping strata may be taken at the boundaries of the forest area.

5.4. Stratified sampling

The basic idea in stratified random sampling is to divide a heterogeneous population into sub-populations, usually known as strata, each of which is internally homogeneous in which case a precise estimate of any stratum mean can be obtained based on a small sample from that stratum and by combining such estimates, a precise estimate for the whole population can be obtained. Stratified sampling provides a better cross section of the population than the procedure of simple random sampling. It may also simplify the organisation of the field work. Geographical proximity is sometimes taken as the basis of stratification. The assumption here is that geographically contiguous areas are often more alike than areas that are far apart. Administrative convenience may also dictate the basis on which the stratification is made. For example, the staff already available in each range of a forest division may have to supervise the survey in the area under their jurisdiction. Thus, compact geographical regions may form the strata. A fairly effective method of stratification is to conduct a quick reconnaissance survey of the area or pool the information already at hand and stratify the forest area according to forest types, stand density, site quality etc. If the characteristic under study is known to be correlated with a supplementary variable for which actual data or at least good estimates are available for the units in the population, the stratification may be done using the information on the supplementary variable. For instance, the volume estimates obtained at a previous inventory of the forest area may be used for stratification of the population.

In stratified sampling, the variance of the estimator consists of only the 'within strata' variation. Thus the larger the number of strata into which a population is divided, the higher, in general, the precision, since it is likely that, in this case, the units within a stratum will be more homogeneous. For estimating the variance within strata, there should be a minimum of 2 units in each stratum. The larger the number of strata the higher will, in general, be the cost of enumeration. So, depending on administrative convenience, cost of the survey and variability of the characteristic under study in the area, a decision on the number of strata will have to be arrived at.

5.4.1. Allocation and selection of the sample within strata

Assume that the population is divided into k strata of N_1, N_2 ,…, N_k units respectively, and that a sample of n units is to be drawn from the population. The problem of allocation concerns the choice of the sample sizes in the respective strata, *i.e.*, how many units should be taken from each stratum such that the total sample is n.

Other things being equal, a larger sample may be taken from a stratum with a larger variance so that the variance of the estimates of strata means gets reduced. The application of the above principle requires advance estimates of the variation within each stratum. These may be available from a previous survey or may be based on pilot

surveys of a restricted nature. Thus if this information is available, the sampling fraction in each stratum may be taken proportional to the standard deviation of each stratum.

In case the cost per unit of conducting the survey in each stratum is known and is varying from stratum to stratum an efficient method of allocation for minimum cost will be to take large samples from the stratum where sampling is cheaper and variability is higher. To apply this procedure one needs information on variability and cost of observation per unit in the different strata.

Where information regarding the relative variances within strata and cost of operations are not available, the allocation in the different strata may be made in proportion to the number of units in them or the total area of each stratum. This method is usually known as 'proportional allocation'.

For the selection of units within strata, In general, any method which is based on a probability selection of units can be adopted. But the selection should be independent in each stratum. If independent random samples are taken from each stratum, the sampling procedure will be known as 'stratified random sampling'. Other modes of selection of sampling such as systematic sampling can also be adopted within the different strata.

5.4.2. Estimation of mean and variance

We shall assume that the population of N units is first divided into k strata of N_1, N_2,…,N_k units respectively. These strata are non-overlapping and together they comprise the whole population, so that

$$N_1 + N_2 + \ldots + N_k = N. \tag{5.21}$$

When the strata have been determined, a sample is drawn from each stratum, the selection being made independently in each stratum. The sample sizes within the strata are denoted by $n_1, n_2, \ldots, n_k$ respectively, so that

$$n_1 + n_2 + \ldots + n_3 = n \tag{5.22}$$

Let y_{tj} (j = 1, 2,…., N_t ; t = 1, 2,..…k) be the value of the characteristic under study for the j the unit in the tth stratum. In this case, the population mean in the tth stratum is given by

$$\overline{Y}_t = \frac{1}{N_t}\sum_{j=1}^{N_t} y_{tj,}(t = 1,\ 2,\ \ldots,\ k) \tag{5.23}$$

The overall population mean is given by

$$\overline{Y} = \frac{1}{N}\sum_{t=1}^{k} N_t \overline{Y}_t \tag{5.24}$$

The estimate of the population mean $\overline{Y}$, in this case will be obtained by

$$\hat{\bar{Y}} = \frac{\sum_{t=1}^{k} N_t \bar{y}_t}{N} \tag{5.25}$$

where $\bar{y}_t = \sum_{j=1}^{n_t} \frac{y_{tj}}{n_t}$ (5.26)

Estimate of the variance of $\hat{\bar{Y}}$ is given by

$$\hat{V}(\hat{\bar{Y}}) = \frac{1}{N^2} \sum_{t=1}^{k} N_t (N_t - n_t) \frac{s^2_{t(y)}}{n_t} \tag{5.27}$$

where $s^2_{t(y)} = \sum_{j=1}^{n_t} \frac{(y_{tj} - \bar{y}_t)^2}{n_t - 1}$ (5.28)

Stratification, if properly done as explained in the previous sections, will usually give lower variance for the estimated population total or mean than a simple random sample of the same size. However, a stratified sample taken without due care and planning may not be better than a simple random sample.

Numerical illustration of calculating the estimate of mean volume per hectare of a particular species and its standard error from a stratified random sample of compartments selected independently with equal probability in each stratum is given below.

A forest area consisting of 69 compartments was divided into three strata containing compartments 1-29, compartments 30-45, and compartments 46 to 69 and 10, 5 and 8 compartments respectively were chosen at random from the three strata. The serial numbers of the selected compartments in each stratum are given in column (4) of Table 5.3. The corresponding observed volume of the particular species in each selected compartment in m^3/ha is shown in column (5).

Table 5.3. Illustration of estimation of parameters under stratified sampling

Stratum number	Total number of units in the stratum (N_t)	Number of units sampled (n_t)	Selected sampling unit number	Volume (m^3/ha) (y_{t_j})	($y^2_{t_j}$)
(1)	(2)	(3)	(4)	(5)	(6)
I			1	5.40	29.16
			18	4.87	23.72
			28	4.61	21.25
			12	3.26	10.63
			20	4.96	24.60
			19	4.73	22.37
			9	4.39	19.27
			6	2.34	5.48
			17	4.74	22.47
			7	2.85	8.12

Total	29	10	..	42.15	187.07
II			43	4.79	22.94
			42	4.57	20.88
			36	4.89	23.91
			45	4.42	19.54
			39	3.44	11.83
Total	16	5	..	22.11	99.10
III			59	7.41	54.91
			50	3.70	13.69
			49	5.45	29.70
			58	7.01	49.14
			54	3.83	14.67
			69	5.25	27.56
			52	4.50	20.25
			47	6.51	42.38
Total	24	8	..	43.66	252.30

Step 1. Compute the following quantities.

$N = (29 + 16 + 24) = 69$

$n = (10 + 5 + 8) = 23$

$\bar{y}_t = 4.215, \quad \bar{y}_t = 4.422, \quad \bar{y}_t = 5.458$

Step 2. Estimate of the population mean $\bar{Y}$ using Equation (3) is

$$\hat{\bar{Y}} = \frac{\sum_{t=1}^{3} N_t \bar{y}_t}{N} = \frac{(29 \times 4.215) + (16 \times 4.422) + (24 \times 5.458)}{69} = \frac{323.979}{69} = 4.70$$

Step 3. Estimate of the variance of $\hat{\bar{Y}}$ using Equation (5) as

$$\hat{V}(\hat{\bar{Y}}) = \frac{1}{N^2} \sum_{t=1}^{3} N_t (N_t - n_t) \frac{s_{t(y)}^2}{n_t}$$

In this example,

$$s_{1(y)}^2 = \frac{187.07 - \frac{(42.15)^2}{10}}{9} = \frac{9.41}{9} = 1.046$$

$$s^2_{2(y)} = \frac{99.10 - \frac{(22.11)^2}{5}}{4} = \frac{1.33}{4} = 0.333$$

$$s^2_{3(y)} = \frac{252.30 - \frac{(43.66)^2}{8}}{7} = \frac{14.03}{7} = 2.004$$

$$\hat{V}(\hat{\bar{Y}}) = \left(\frac{1}{69}\right)^2\left[\left(\frac{29\times19}{10}\times1.046\right)+\left(\frac{16\times11}{5}\times0.333\right)+\left(\frac{24\times16}{8}\times2.004\right)\right]$$

$$= \frac{165.5482}{4761} = 0.03477$$

$$SE(\hat{\bar{Y}}) = \sqrt{0.03477} = 0.1865$$

$$RSE(\hat{\bar{Y}}) = \frac{SE(\hat{\bar{Y}})\times100}{\hat{\bar{Y}}} \qquad (5.29)$$

$$= \frac{0.1865\times100}{4.70} = 3.97\%$$

Now, if we ignore the strata and assume that the same sample of size n = 23, formed a simple random sample from the population of N = 69, the estimate of the population mean would reduce to

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{42.15 + 22.11 + 43.66}{23} = \frac{107.92}{23} = 4.69$$

Estimate of the variance of the mean $\bar{y}$ is

$$\hat{V}(\bar{y}) = \frac{N-n}{Nn}s^2$$

where

$$s^2 = \frac{538.47 - \frac{(107.92)^2}{23}}{22}$$

$$= \frac{32.09}{22} = 1.4586$$

so that

$$\hat{V}(\bar{y}) = \frac{(69-23)}{69\times23}\times1.4586$$

$$= \frac{2.9172}{69} = 0.04230$$

$$SE(\bar{y}) = \sqrt{0.04230} = 0.2057$$

$$RSE(\bar{y}) = \frac{0.2057 \times 100}{4.69} = 4.39\%$$

The gain in precision due to stratification is computed by

$$\frac{\hat{V}(\hat{\bar{Y}})_{srs}}{\hat{V}(\hat{\bar{Y}})_{st}} \times 100 = \frac{0.04230}{0.03477} \times 100$$
$$= 121.8$$

Thus the gain in precision is 21.8%.

5.5. Multistage sampling

With a view to reduce cost and/or to concentrate the field operations around selected points and at the same time obtain precise estimates, sampling is sometimes carried out in stages. The procedure of first selecting large sized units and then choosing a specified number of sub-units from the selected large units is known as sub-sampling. The large units are called 'first stage units' and the sub-units the 'second stage units'. The procedure can be easily generalised to three stage or multistage samples. For example, the sampling of a forest area may be done in three stages, firstly by selecting a sample of compartments as first stage units, secondly, by choosing a sample of topographical sections in each selected compartment and lastly, by taking a number of sample plots of a specified size and shape in each selected topographical section.

The multistage sampling scheme has the advantage of concentrating the sample around several 'sample points' rather than spreading it over the entire area to be surveyed. This reduces considerably the cost of operations of the survey and helps to reduce the non-sampling errors by efficient supervision. Moreover, in forest surveys it often happens that detailed information may be easily available only for groups of sampling units but not for individual units. Thus, for example, a list of compartments with details of area may be available but the details of the topographical sections in each compartment may not be available. Hence if compartments are selected as first stage units, it may be practicable to collect details regarding the topographical sections for selected compartments only and thus use a two-stage sampling scheme without attempting to make a frame of the topographical sections in all compartments. The multistage sampling scheme, thus, enables one to use an incomplete sampling frame of all the sampling units and to properly utilise the information already available at every stage in an efficient manner.

The selection at each stage, in general may be either simple random or any other probability sampling method and the method may be different at the different stages. For example one may select a simple random sample of compartments and take a systematic line plot survey or strip survey with a random start in the selected compartments.

5.5.1. Two-stage simple random sampling

When at both stages the selection is by simple random sampling, method is known as two stage simple random sampling. For example, in estimating the weight of grass in a forest area, consisting of 40 compartments, the compartments may be considered as primary sampling units. Out of these 40 compartments, n = 8 compartments may be selected randomly using simple random sampling procedure as illustrated in Section 5.2.1. A random sample of plots either equal or unequal in number may be selected from each selected compartment for the measurement of the quantity of grass through the procedure of selecting a simple random sample. It is then possible to develop estimates of either mean or total quantity of grass available in the forest area through appropriate formulae.

5.5.2. Parameter estimation under two-stage simple random sampling

Let the population consists of N first stage units and let M_i be the number of second stage units in the ith first stage unit. Let n first stage units be selected and from the ith selected first stage unit let m_i second stage units be chosen to form a sample of $m = \sum_{i=1}^{n} m_i$ units. Let y_{ij} be the value of the character for the jth second stage unit in the ith first stage unit.

An unbiased estimator of the population mean $\overline{Y} = \dfrac{\sum_{i}^{N}\sum_{j}^{M_i} y_{ij}}{\sum_{i=1}^{N} M_i}$ is obtained by Equation (5.30).

$$\hat{\overline{Y}} = \frac{1}{n\overline{M}} \sum_{i=1}^{n} \frac{M_i}{m_i} \sum_{j=1}^{m_i} y_{ij} \qquad (5.30)$$

where $\overline{M} = \dfrac{\sum_{i=1}^{N} M_i}{N}$. (5.31)

The estimate of the variance of $\hat{\overline{Y}}$ is given by

$$\hat{V}(\hat{\overline{Y}}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_b^2 + \frac{1}{nN} \sum_{i=1}^{n} \left(\frac{M_i}{\overline{M}}\right)^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_{w_i}^2 \qquad (5.32)$$

where $s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{M_i}{\overline{M}}\bar{y}_i - \bar{y}\right)^2$ (5.33)

$$s_{w_i}^2 = \frac{1}{m_i - 1}\sum_{j=1}^{m_i}(y_{ij} - \bar{y}_i)^2 \quad (5.34)$$

The variance of $\hat{\bar{Y}}$ here can be noticed to be composed of two components. The first is a measure of variation between first stage units and the second, a measure of variation within first stage units. If $m_i = M_i$, the variance is given by the first component only. The second term, thus represents the contribution due to sub-sampling.

An example of the analysis of a two stage sample is given below. Table 5.4 gives data on weight of grass (mixed species) in kg from plots of size 0.025 ha selected from 8 compartments which were selected randomly out of 40 compartments from a forest area. The total forest area was 1800ha.

Table 5.4. Weight of grass in kg in plots selected through a two stage sampling procedure.

Plot	Compartment number								Total
	I	II	III	IV	V	VI	VII	VIII	
1	96	98	135	142	118	80	76	110	
2	100	142	88	130	95	73	62	125	
3	113	143	87	106	109	96	105	77	
4	112	84	108	96	147	113	125	62	
5	88	89	145	91	91	125	99	70	
6	139	90	129	88	125	68	64	98	
7	140	89	84	99	115	130	135	65	
8	143	94	96	140	132	76	78	97	
9	131	125	..	98	148	84	..	106	
10	..	116	..	..	..	105	..	..	
Total	1062	1070	872	990	1080	950	744	810	7578
m_i	9	10	8	9	9	10	8	9	72
Mean ($\bar{y}_i$)	118	107	109	110	120	95	93	90	842
M_i	1760	1975	1615	1785	1775	2050	1680	1865	14505
$s_{w_i}^2$	436.00	515.78	584.57	455.75	412.25	496.67	754.86	496.50	4152
$\frac{s_{w_i}^2}{m_i}$	48.44	51.578	73.07	50.63	45.80	49.667	94.35	55.167	

Step1. Estimate the mean weight of grass in kg per plot using the formula in Equation (5.30).

$$\hat{\bar{Y}} = \frac{1}{n\bar{M}} \sum_{i=1}^{n} \frac{M_i}{m_i} \sum_{j=1}^{m_i} y_{ij}$$

$$\bar{M} = \frac{1}{N} \sum_{i=1}^{N} M_i = \frac{1}{40} \left(\frac{1800}{0.025} \right)$$

$$= 1800$$

Since $\sum M_i$ indicates the total number of second stage units, it can be obtained by dividing the total area (1800 ha) by the size of a second stage unit (0.025 ha).

Estimate of the population mean calculated using Equation (5.30) is

$$\hat{\bar{Y}} = \frac{1}{n\bar{M}} \sum_{i=1}^{n} \frac{M_i}{m_i} \sum_{j=1}^{m_i} y_{ij}$$

$$= \frac{1523230}{(8)(40)} = 105.78$$

$$s_b^2 = \frac{1}{(8-1)} \left[\left(\frac{1760}{1800} \times 118 - 105.25 \right)^2 + \left(\frac{1975}{1800} \times 107 - 105.25 \right)^2 + \ldots + \left(\frac{1865}{1800} \times 90 - 105.25 \right)^2 \right]$$

$$=140.36$$

Estimate of variance of $\hat{\bar{Y}}$ obtained by Equation (5.32) is

$$\hat{V}(\hat{\bar{Y}}) = \left(\frac{1}{8} - \frac{1}{40} \right) 140.3572 + \frac{1}{(8)(40)} (465.1024)$$

$$=15.4892$$

$$SE(\hat{\bar{Y}}) = \sqrt{15.4892} = 3.9356$$

$$RSE(\hat{\bar{Y}}) = \frac{3.9356 \times 100}{105.78} = 3.72\%$$

5.6. Multiphase sampling

Multiphase sampling plays a vital role in forest surveys with its application extending over continuous forest inventory to estimation of growing stock through remote

sensing. The essential idea in multiphase sampling is that of conducting separate sampling investigations in a sequence of phases starting with a large number of sampling units in the first phase and taking only a subset of the sampling units in each successive phase for measurement so as to estimate the parameter of interest with added precision at relatively lower cost utilizing the relation between characters measured at different phases. In order to keep things simple, further discussion in this section is restricted to only two phase sampling.

A sampling technique which involves sampling in just two phases (occasions) is known as two phase sampling. This technique is also referred to as double sampling. Double sampling is particularly useful in situations in which the enumeration of the character under study (main character) involves much cost or labour whereas an auxiliary character correlated with the main character can be easily observed. Thus it can be convenient and economical to take a large sample for the auxiliary variable in the first phase leading to precise estimates of the population total or mean of the auxiliary variable. In the second phase, a small sample, usually a sub-sample, is taken wherein both the main character and the auxiliary character may be observed and using the first phase sampling as supplementary information and utilising the ratio or regression estimates, precise estimates for the main character can be obtained. It may be also possible to increase the precision of the final estimates by including instead of one, a number of correlated auxiliary variables. For example, in estimating the volume of a stand, we may use diameter or girth of trees and height as auxiliary variables. In estimating the yield of tannin materials from bark of trees certain physical measurements like the girth, height, number of shoots, etc., can be taken as auxiliary variables.

Like many other kinds of sampling, double sampling is a technique useful in reducing the cost of enumerations and increasing the accuracy of the estimates. This technique can be used very advantageously in resurveys of forest areas. After an initial survey of an area, the estimate of growing stock at a subsequent, period, say 10 or 15 years later, and estimate of the change in growing stock can be obtained based on a relatively small sample using double sampling technique.

Another use of double sampling is in stratification of a population. A first stage sample for an auxiliary character may be used to sub-divide the population into strata in which the second (main) character varies little so that if the two characters are correlated, precise estimates of the main character can be obtained from a rather small second sample for the main character.

It may be mentioned that it is possible to couple with double sampling other methods of sampling like multistage sampling (sub-sampling)known for economy and enhancing the accuracy of the estimates. For example, in estimating the availability of grasses, canes, reeds, etc., a two-stage sample of compartments (or ranges) and topographical sections (or blocks) may be taken for the estimation of the effective area under the species and a sub-sample of topographical sections, blocks or plots may be taken for estimating the yield.

5.6.1. Selection of sampling units

In the simplest case of two phase sampling, simple random sampling can be employed in both the phases. In the first step, the population is divided into well identified sampling units and a sample is drawn as in the case of simple random sampling. The character x is measured on all the sampling units thus selected. Next, a sub-sample is taken from the already selected units using the method of simple random sampling and the main character of interest (y) is measured on the units selected. The whole procedure can also be executed in combination with other modes of sampling such as stratification or multistage sampling schemes.

5.6.2. Parameter estimation

(i) *Regression estimate in double sampling :*

Let us assume that a random sample of n units has been taken from the population of N units at the initial phase to observe the auxiliary variable x and that a random sub-sample of size m is taken where both x and the main character y are observed.

Let $\bar{x}_{(n)}$ = mean of x in the first large sample = $\bar{x}_{(n)} = \sum_{i=1}^{n} \frac{x_i}{n}$ (5.35)

$\bar{x}_{(m)}$ = mean of x in the second sample = $\bar{x}_{(m)} = \sum_{i=1}^{m} \frac{x_i}{m}$ (5.36)

$\bar{y}$ = mean of y in the second sample = $\bar{y} = \sum_{i=1}^{m} \frac{y_i}{m}$ (5.37)

We may take $\bar{y}$ as an estimate of the population mean $\bar{Y}$. However utilising the previous information on the units sampled, a more precise estimate of $\bar{Y}$ can be obtained by calculating the regression of y on x and using the first sample as providing supplementary information. The regression estimate of $\bar{Y}$ is given by

$$\bar{y}_{(drg)} = \bar{y} + b(\bar{x}_{(n)} - \bar{x}_{(m)}) \tag{5.38}$$

where the suffix *(drg)* denotes the regression estimate using double sampling and b is the regression coefficient of y on x computed from the units included in the second sample of size m. Thus

$$b = \frac{\sum_{i=1}^{m}\left(x_i - \bar{x}_{(m)}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{m}\left(x_i - \bar{x}_{(m)}\right)^2} \tag{5.39}$$

The variance of the estimate is approximately given by,

$$V(\bar{y})_{(drg)} = \frac{s^2_{y.x}}{m} + \frac{s^2_{y.x} - s^2_y}{n} \tag{5.40}$$

where $$s^2_{y.x} = \frac{1}{m-2}\left[\sum_{i=1}^{m}(y_i - \bar{y})^2 - b^2\sum_{i=1}^{m}(x_i - \bar{x}_{(m)})^2\right] \tag{5.41}$$

$$s^2_y = \sum_{i=1}^{m}\frac{(y_i - \bar{y})^2}{m-1} \tag{5.42}$$

(ii) Ratio estimate in double sampling

Ratio estimate is used mainly when the intercept in the regression line between y and x is understood to be zero. The ratio estimate of the population mean $\bar{Y}$ is given by

$$\bar{y}_{(dra)} = \frac{\bar{y}}{\bar{x}_{(m)}}\bar{x}_{(n)} \tag{5.43}$$

where $\bar{y}_{dra}$ denotes the ratio estimate using double sampling. The variance of the estimate is approximately given by

$$V(\bar{y}_{dra}) = \frac{s^2_y - 2\hat{R}s_{yx} + \hat{R}^2 s^2_x}{m} + \frac{2\hat{R}s_{yx} - \hat{R}^2 s^2_x}{n} \tag{5.44}$$

where

$$s^2_y = \frac{\sum_{i=1}^{m}(y_i - \bar{y})^2}{m-1} \tag{5.45}$$

$$s_{yx} = \frac{\sum_{i=1}^{m}(y_i - \bar{y})(x_i - \bar{x}_{(m)})}{m-1} \tag{5.46}$$

$$s^2_x = \frac{\sum_{i-1}^{m}(x_i - \bar{x}_m)^2}{m-1} \tag{5.47}$$

$$\hat{R} = \frac{\bar{y}}{\bar{x}_{(m)}} \tag{5.48}$$

An example of analysis of data from double sampling using regression and ratio estimate is given below. Table 5.5 gives data on the number of clumps and the corresponding weight of grass in plots of size 0.025 ha, obtained from a random sub-sample of 40 plots taken from a preliminary sample of 200 plots where only the number of clumps was counted.

Table 5.5. Data on the number of clumps and weight of grass in plots selected through a two phase sampling procedure.

Serial number	Number of clumps	Weight in kg	Serial number	Number of clumps	Weight in kg

	(x)	(y)		(x)	(y)
1	459	68	21	245	25
2	388	65	22	185	50
3	314	44	23	59	16
4	35	15	24	114	22
5	120	34	25	354	59
6	136	30	26	476	63
7	367	54	27	818	92
8	568	69	28	709	64
9	764	72	29	526	72
10	607	65	30	329	46
11	886	95	31	169	33
12	507	60	32	648	74
13	417	72	33	446	61
14	389	60	34	86	32
15	258	50	35	191	35
16	214	30	36	342	40
17	674	70	37	227	40
18	395	57	38	462	66
19	260	45	39	592	68
20	281	36	40	402	55

Here, $n = 200$, $m = 40$. The mean number of clumps per plot as observed from the preliminary sample of 200 plots was $\bar{x}_{(n)} = 374.4$.

$$\sum_{i=1}^{40} x_i = 15419, \quad \sum_{i=1}^{40} y_i = 2104,$$

$$\sum_{i=1}^{40} x_i^2 = 7744481, \quad \sum_{i=1}^{40} y_i^2 = 125346, \quad \sum_{i=1}^{40} x_i y_i = 960320$$

$$\sum_{1}^{40} (x_i - \bar{x}_{(m)})^2 = \sum_{i=1}^{40} x_i^2 - \frac{\left(\sum_{1}^{40} x_i\right)^2}{40} = 7744481 - \frac{\left(15419^2\right)}{40} = 1800842$$

$$\sum_{1}^{40} \left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{40} y_i^2 - \frac{\left(\sum_{i=1}^{40} y_i\right)^2}{40} = 125346 - \frac{(2104)^2}{40} = 14675.6$$

$$\sum_{1}^{40} \left(x_i - \bar{x}_{(m)}\right)\left(y_i - \bar{y}\right) = \sum_{1}^{40} x_i y_i - \frac{\sum_{1}^{40} x_i \sum_{1}^{40} y_i}{40} = 960320 - \frac{15419 \times 2104}{40} = 149280.6$$

Mean number of clumps per plot from the sub-sample of 40 plots is

$$\bar{x}_{(m)} = \frac{15419}{40} = 385.5$$

Mean weight of clumps per plot from the sub-sample of 40 plots

$$\bar{y} = \frac{2104}{40} = 52.6$$

The regression estimate of the mean weight of grass in kg per plot is obtained by using Equation (5.38) where the regression coefficient *b* obtained using Equation (5.39) is

$$b = \frac{149280.6}{1800842} = 0.08$$

Hence, $\bar{y}_{(drg)} = 52.6 + 0.08(374.4 - 385.5)$

= 52.6 - 0.89

= 51.7 kg /plot

$$s^2_{y.x} = \frac{1}{40-2}\left[14675.6 - (0.08)^2(1800842)\right]$$

= 82.9

$$s^2_y = \frac{14675.6}{39}$$

=376.297

The variance of the estimate is approximately given by Equation (5.40)

$$V(\bar{y})_{(drg)} = \frac{82.9}{40} + \frac{82.9 - 376.297}{200} \qquad (5.40)$$

= 3.5395

The ratio estimate of the mean weight of grass in kg per plot is given by Equation (5.43)

$$\bar{y}_{(dra)} = \frac{52.6}{385.5}(374.4)$$

= 51.085

$$s_{yx} = \frac{149280.6}{40-1}$$

= 3827.708

$$s_x^2 = \frac{1800842}{40-1}$$
$$= 46175.436$$

$$\hat{R} = \frac{52.6}{385.5}$$
$$= 0.1364$$

The variance of the estimate is approximately given by Equation (5.44) is

$$V(\bar{y}_{dra}) = \frac{376.297 - 2(0.1364)(3827.708) + (0.1364)^2(46175.436)}{40}$$
$$+ \frac{(2)(0.1364)(3827.708) - (0.1364)^2(46175.436)}{200}$$
$$= 5.67$$

5.7. Probability Proportional to Size (PPS) sampling

In many instances, the sampling units vary considerably in size and simple random sampling may not be effective in such cases as it does not take into account the possible importance of the larger units in the population. In such cases, it has been found that ancillary information about the size of the units can be gainfully utilised in selecting the sample so as to get a more efficient estimator of the population parameters. One such method is to assign unequal probabilities for selection to different units of the population. For example, villages with larger geographical area are likely to have larger area under food crops and in estimating the production, it would be desirable to adopt a sampling scheme in which villages are selected with probability proportional to geographical area. When units vary in their size and the variable under study is directly related with the size of the unit, the probabilities may be assigned proportional to the size of the unit. This type of sampling where the probability of selection is proportion to the size of the unit is known as 'PPS Sampling'. While sampling successive units from the population, the units already selected can be replaced back in the population or not. In the following, PPS sampling with replacement of sampling units is discussed as this scheme is simpler compared to the latter.

5.7.1. Methods of selecting a pps sample with replacement

The procedure of selecting the sample consists in associating with each unit a number or numbers equal to its size and selecting the unit corresponding to a number chosen at random from the totality of numbers associated. There are two methods of selection which are discussed below:

(i) Cumulative total method: Let the size of the *i*th unit be x_i, (i = 1, 2, …, N). We associate the numbers 1 to x_i with the first unit, the numbers (x_1+1) to (x_1+x_2) with the

second unit and so on such that the total of the numbers so associated is $X = x_1 + x_2 + \ldots + x_N$. Then a random number r is chosen at random from 1 to X and the unit with which this number is associated is selected.

For example, a village has 8 orchards containing 50, 30, 25, 40, 26, 44, 20 and 35 trees respectively. A sample of 3 orchards has to be selected with replacement and with probability proportional to number of trees in the orchards. We prepare the following cumulative total table:

Serial number of the orchard	Size (x_i)	Cumulative size	Numbers associated
1	50	50	1 - 50
2	30	80	51 - 80
3	25	105	81 -105
4	40	145	106 -145
5	26	171	146 - 171
6	44	215	172 - 215
7	20	235	216 - 235
8	35	270	236 - 270

Now, we select three random numbers between 1 and 270. The random numbers selected are 200, 116 and 47. The units associated with these three numbers are 6^{th}, 4^{th}, and 1^{st} respectively. And hence, the sample so selected contains units with serial numbers, 1, 4 and 6.

(ii) Lahiri's Method: We have noticed that the cumulative total method involves writing down the successive cumulative totals which is time consuming and tedious, especially with large populations. Lahiri in 1951 suggested an alternative procedure which avoids the necessity of writing down the cumulative totals. Lahiri's method consists in selecting a pair of random numbers, say (i,j) such that $1 \leq i \leq N$ and $1 \leq j \leq M$; where M is the maximum of the sizes of the N units of the population. If $j \leq X_i$, the ith unit is selected: otherwise, the pair of random number is rejected and another pair is chosen. For selecting a sample of n units, the procedure is to be repeated till n units are selected. This procedure leads to the required probabilities of selection.

For instance, to select a sample of 3 orchards from the population in the previous example in this section, by Lahiri's method by PPS with replacement, as $N = 8$, $M = 50$ and $n = 3$, we have to select three pairs of random numbers such that the first random number is less than or equal to 8 and the second random number is less than or equal to 50. Referring to the random number table, three pairs selected are (2, 23) (7,8) and (3, 30). As in the third pair $j > X_i$, a fresh pair has to be selected. The next pair of random numbers from the same table is (2, 18) and hence, the sample so selected consists of the units with serial numbers 2, 7 and 2. Since the sampling unit 2 gets repeated in the sample, the effective sample size is two in this case. In order to get an effective sample size of three, one may repeat the sampling procedure to get another distinct unit.

5.7.2. Estimation procedure

Let a sample of n units be drawn from a population consisting of N units by PPS with replacement. Further, let (y_i, p_i) be the value and the probability of selection of the ith unit of the sample, $i = 1, 2, 3, \ldots, n$.

An unbiased estimator of population mean is given by

$$\hat{\bar{Y}} = \frac{1}{nN}\sum_{i=1}^{n}\frac{y_i}{p_i} \tag{5.49}$$

An estimator of the variance of above estimator is given by

$$\hat{V}\left(\hat{\bar{Y}}\right) = \frac{1}{n(n-1\_N^2}\left(\sum_{i=1}^{N}\left(\frac{y_i}{p_i}\right)^2 - n\hat{Y}^2\right) \tag{5.50}$$

where $p_i = \frac{x_i}{X}$, $\hat{Y} = N\hat{\bar{Y}}$

For illustration, consider the following example. A random sample 23 units out of 69 units were selected with probability proportional to size of the unit (compartment) from a forest area in U.P. The total area of 69 units was 14079 ha. The volume of timber determined for each selected compartment are given in Table 5.6 along with the area of the compartment.

Table 5. 6. Volume of timber and size of the sampling unit for a PPS sample of forest compartments.

Serial no.	Size in ha (x_i)	Relative size (x_i/X)	Volume in m^3 (y_i)	$\frac{y_i}{p_i} = v_i$	$(v_i)^2$
1	135	0.0096	608	63407.644	4020529373.993
2	368	0.0261	3263	124836.351	15584114417.014
3	374	0.0266	877	33014.126	1089932493.652
4	303	0.0215	1824	84752.792	7183035765.221
5	198	0.0141	819	58235.864	3391415813.473
6	152	0.0108	495	45849.375	2102165187.891
7	264	0.0188	1249	66608.602	4436705896.726
8	235	0.0167	1093	65482.328	4287935235.716
9	467	0.0332	1432	43171.580	1863785345.581
10	458	0.0325	3045	93603.832	8761677342.194
11	144	0.0102	410	40086.042	1606890736.502
12	210	0.0149	1460	97882.571	9580997789.469
13	467	0.0332	1432	43171.580	1863785345.581
14	458	0.0325	3045	93603.832	8761677342.194
15	184	0.0131	1003	76745.853	5889925992.739
16	174	0.0124	834	67482.103	4553834285.804
17	184	0.0131	1003	76745.853	5889925992.739
18	285	0.0202	2852	140888.800	19849653965.440
19	621	0.0441	4528	102656.541	10538365422.979

20	111	0.0079	632	80161.514	6425868248.777
21	374	0.0266	877	33014.126	1089932493.652
22	64	0.0045	589	129570.797	16788591402.823
23	516	0.0367	1553	42373.424	1795507096.959
				1703345.530	147356252987.120

Total area X = 14079 ha.

An unbiased estimator of population mean is obtained by using Equation (5.49).

$$\hat{\bar{Y}} = \frac{1}{(23)(69)}(1703345.530)$$
$$= 1073.312$$

An estimate of the variance of $\hat{\bar{Y}}$ is obtained through Equation (5.50).

$$\hat{V}\left(\hat{\bar{Y}}\right) = \frac{1}{23(23-1)(69)^2}(147356252987.120 - (23)(67618.632))$$
$$= 17514.6$$

And the standard error of $\bar{Y}$ is $\sqrt{17514.6}$ = 132.343.

6. SPECIAL TOPICS

A number of instances in forestry research can be found wherein substantial statistical applications have been made other than the regular design, sampling or analytical techniques. These special methods are integrally related to the concepts in the particular subject fields and will require an understanding of both statistics and the concerned disciplines to fully appreciate their implications. Some of these topics are briefly covered in what follows. It may be noted that quite many developments have taken place in each of the topics mentioned below and what is reported here forms only a basic set in this respect. The reader is prompted to make further reading wherever required so as to get a better understanding of the variations possible with respect to data structure or in the form of analysis in such cases.

6.1 Genetics and plant breeding

6.1.1. Estimation of heritability and genetic gain

The observed variation in a group of individuals is partly composed of genetic or heritable variation and partly of non-heritable variation. The fraction of total variation which is heritable is termed the coefficient of heritability in the broad sense. The genotypic variation itself can be sub-divided into additive and nonadditive genetic

variance. The ratio of additive genetic variance to the total phenotypic variance is called the coefficient of heritability in the narrow sense and is designated by h^2. Thus,

$$h^2 = \frac{\text{additive genetic variance}}{\text{additive genetic variance} + \text{nonadditive genetic variance} + \text{environmental variance}}$$

Conceptually, genetic gain or genetic improvement per generation is the increase in productivity following a change in the gene frequency induced mostly by selection.

Heritability and genetic gain can be estimated in either of two ways. The most direct estimates are derived from the relation between parents and offspring, obtained by measuring the parents, growing the offspring, and measuring the offspring. The other way is to establish a half-sib or full-sib progeny test, conduct an analysis of variance and compute heritability as a function of the variances. Understanding the theoretical part in this context requires a thorough knowledge of statistics. Formulae given below in this section are intended only as handy references. Also, there is no attempt made to cover the numerous possible variations that might result from irregularities in design. Half-sib progeny test is used for illustration as it is easier to establish and so more common in forestry.

Both heritability and gain estimates apply strictly only to the experiments from which they are obtained. They may be and frequently are very different when obtained from slightly different experiments. Therefore when quoting them, it is desirable to include pertinent details of experimental design and calculation procedures. Also, it is good practice to state the statistical reliability of each heritability estimate and therefore formulae for calculating the reliability of heritability estimates are also included in this section. Additional references are Falconer (1960), Jain (1982) and Namkoong *et al.* (1966).

For illustration of the techniques involved, consider the data given in Table 6.1. The data were obtained from a replicated progeny trial in bamboo conducted at Vellanikkara and Nilambur in Kerala consisting of 6 families, replicated 3 times at each of the 2 sites, using plots of 6 trees each. The data shown in Table 6.1 formed part of a larger set.

Table 6.1. Data on height obtained from a replicated progeny trial in bamboo conducted at 2 sites in Kerala.

		Height (cm) after two years of planting											
		Site I - Vellanikkara						Site II - Nilambur					
		Family						Family					
Block	Tree	1	2	3	4	5	6	1	2	3	4	5	6
1	1	142	104	152	111	23	153	24	18	18	31	95	57

	2	95	77	98	29	48	51	58	50	24	26	42	94
	3	138	129	85	64	88	181	32	82	38	30	43	77
	4	53	126	118	52	27	212	27	23	65	86	76	39
	5	95	68	25	19	26	161	60	56	46	20	41	82
	6	128	48	51	25	26	210	75	61	104	28	49	29
2	1	185	129	78	28	35	140	87	26	78	25	29	54
	2	117	131	161	26	21	79	102	103	57	37	72	56
	3	135	135	121	25	14	158	74	55	60	52	83	29
	4	155	88	124	76	34	93	102	43	26	139	40	67
	5	152	75	118	43	49	151	20	100	59	49	24	42
	6	111	41	61	86	31	171	80	98	70	97	54	47
3	1	134	53	145	53	72	109	54	58	87	17	25	38
	2	35	82	86	32	113	50	92	47	93	23	30	38
	3	128	71	141	24	37	64	89	33	70	29	26	36
	4	89	43	156	182	19	82	144	108	47	30	36	72
	5	99	71	121	22	24	77	100	70	26	87	24	106
	6	29	26	55	52	20	123	92	46	40	31	37	61

The stepwise procedure for estimating heritability and genetic gain from a half-sib progeny trial is given below.

Step 1. Establish a replicated progeny test consisting of open-pollinated offspring of f families, replicated b (for block) times at each of s sites, using n-tree plots. Measure a trait, such as height, and calculate the analysis of variance, as shown in Table 6.2. Progeny arising from any particular female constitute a family.

Table 6.2. Schematic representation of analysis of variance for a multi-plantation half-sib progeny trial.

Source of variation	Degree of freedom (df)	Sum of squares (SS)	Mean square $\left(MS = \frac{SS}{df}\right)$
Site	s - 1	*SSS*	*MSS*
Block-within-site	s (b - 1)	*SSB*	*MSB*
Family	f - 1	*SSF*	*MSF*
Family x Site	(f - 1)(s - 1)	*SSFS*	*MSFS*
Family x Block-within-site	s(f - 1) (b - 1)	*SSFB*	*MSFB*
Tree-within-plot	*bsf* (n - 1)	*SSR*	*MSR*

The formulae for computing the different sums of squares in the ANOVA Table are given below, including the formula for computing the correction factor

(*C.F.*). Let y_{ijkl} represent the observation corresponding to the *l*th tree belonging to the *k*th family of the *j*th block in the *i*th site. Let *G* represent the grand total, S_i indicate the *i*th site total, F_k represent the *k*th family total, $(SB)_{ij}$ represent the *j*th block total in the *i*th site, $(SF)_{ik}$ represent the *k*th family total in the *i*th site, $(SBF)_{ijk}$ represent the *k*th family total in the *j*th block of the *i*th site.

$$C\,F = \frac{G^2}{sbfn} \tag{6.1}$$

$$= \frac{15418.00^2}{(2)(3)(6)(6)}$$

$$= 1100531.13$$

$$SSTO = \sum_{i=1}^{s}\sum_{j=1}^{b}\sum_{k=1}^{f}\sum_{l=1}^{n} y_{ijkl}^2 - C.F. \tag{6.2}$$

$$= (142)^2+(95)^2+\ldots\ldots+(61)^2 - 1100531.13$$

$$= 408024.87$$

$$SSS = \frac{\sum_{i=1}^{s} S_i^2}{bfn} - C.F. \tag{6.3}$$

$$= \frac{(9334.00)^2 + (6084.00)^2}{(3)(6)(6)} - 1100531.13$$

$$= 48900.46$$

$$SSB = \frac{\sum_{i=1}^{s}\sum_{j=1}^{b}(SB)_{ij}^2}{fn} - C.\,F. - SSS \tag{6.4}$$

$$= \frac{(3238.00)^2 + (3377.00)^2 + \ldots\ldots + (2042.00)^2}{(6)(6)} - 1100531.13 - 48900.46$$

$$= 9258.13$$

$$SSF = \frac{\sum_{k=1}^{f} F_k^2}{sbn} - C.F. \tag{6.5}$$

$$= \frac{(3332.00)^2 + (2574.00)^2 + \ldots\ldots + (3289.00)^2}{(2)(3)(6)} - 1100531.13$$

$$= 80533.37$$

$$SSFS = \frac{\sum_{i=1}^{s}\sum_{j=1}^{b}(SF)_{ik}^2}{bn} - C.\,F. - SSS - SSF \qquad (6.6)$$

$$= \frac{(2020.00)^2 + (1497.00)^2 + \ldots (1024.00)^2}{(3)(6)} - 1100531.13 - 48900.46 - 80533.37$$

$$= 35349.37$$

$$SSFB = \frac{\sum_{i=1}^{s}\sum_{j=1}^{b}\sum_{k=1}^{f}(SBF)_{ijk}^2}{n} - C.\,F. - SSS - SSB - SSF - SSFS \qquad (6.7)$$

$$= \frac{(651.00)^2 + (552.00)^2 + \ldots + (351.00)^2}{(6)} - 1100531.13 - 48900.46 - 9258.13 - 80533.37 - 35349.37$$

$$= 45183.87$$

$$SSR = SSTO - SSS - SSB - SSF - SSFS - SSFB \qquad (6.8)$$

$$= 408024.87 - 48900.46 - 9258.13 - 80533.37 - 35349.37 - 45183.87$$
$$= 188799.67$$

The mean squares are computed as usual by dividing the sums of squares by the respective degrees of freedom. The above results may be summarised as shown in Table 6.3.

Table 6.3. Analysis of variance table for a multi-plantation half-sib progeny trial using data given in Table 6.1.

Source of variation	Degree of freedom (*df*)	Sum of squares (*SS*)	Mean square $\left(MS = \frac{SS}{df}\right)$
Site	1	48900.46	48900.46
Block-within-site	4	9258.13	2314.53
Family	5	80533.37	16106.67
Family x Site	5	35349.37	7069.87
Family x Block-within-site	20	45183.87	2259.19
Tree-within-plot	180	188799.67	1048.89

In ordinary statistical work, the mean squares are divided by each other in various manners to obtain *F* values, which are then used to test the significance. The mean squares themselves, however, are complex, most of them containing variability due to several factors. To eliminate this difficulty, mean squares are

apportioned into variance components according to the equivalents shown in Table 6.4.

Table 6.4. Variance components of mean squares for a multi-plantation half-sib progeny test.

Sources of variation	Variance components of mean squares
Site	$V_e + n\,V_{fb} + n\,b\,V_{fs} + nf\,V_b + nfb\,V_s$
Block-within-site	$V_e + n\,V_{fb} + nf\,V_b$
Family	$V_e + n\,V_{fb} + n\,b\,V_{fs} + nbs\,V_f$
Family x Site	$V_e + n\,V_{fb} + nb\,V_{fs}$
Family x Block-within-site	$V_e + n\,V_{fb}$
Tree-within-plot	V_e

In Table 6.4, V_e , V_{fb} , V_{fs} , V_f , V_b , and V_s are the variances due to tree-within-plot, family x block-within-site, family x site, family, block-within-site and site, respectively.

Step 2. Having calculated the mean squares, set each equal to its variance component as shown in Table 6.4. Start at the bottom of the table and obtain each successive variance of interest by a process of subtraction and division. That is, subtract within-plot mean square (V_e) from family x block mean square ($V_e + nsV_{fb}$) to obtain nsV_{fb} ; then divide by ns to obtain V_{fb} . Proceed in a similar manner up the table.

Step 3.Having calculated the variances, calculate heritability of the half-sib family averages as follows.

$$\text{Family heritability} = \frac{V_f}{\frac{V_e}{nbs} + \frac{V_{fb}}{bs} + \frac{V_{fs}}{s} + V_f} \qquad (6.9)$$

$$= \frac{251.02}{\frac{1048.89}{(6)(3)(2)} + \frac{201.72}{(3)(2)} + \frac{267.26}{(2)} + 251.02}$$

$$= 0.1600$$

Since the family averages are more reliable than the averages for any single plot or tree, the selection is usually based upon family averages.

Step 4. In case the selection is based on the performance of single trees, then single tree heritability is to be calculated. In a half-sib progeny test, differences among families account for only one-fourth of the additive genetic variance; the remainder is accounted for by variation within the families. For that reason, V_f is multiplied by 4 when calculating single tree heritability. Also, since selection is

based upon single trees, all variances are inserted *in toto* in the denominator. Therefore the formula for single-tree heritability is

$$\text{Single tree heritability} = \frac{4V_f}{V_e + V_{fb} + V_{fs} + V_f} \qquad (6.10)$$

$$= \frac{(4)(251.02)}{1048.89 + 201.72 + 267.26 + 251.02}$$

$$= 0.5676$$

Suppose that the families are tested in only one test plantation. Testing and calculating procedure are much simplified. Total degrees of freedom are *nfb* -1; site and family x site mean squares and variances are eliminated from Table 6.2. In this situation, families are measured at one site only. They might grow very differently at other sites. The calculated V_f is in reality a combination of V_f and V_{fs}. Therefore, heritability calculated on the basis of data from one plantation only, is overestimated.

Recording and analysis of single tree data are the most laborious parts of measurement and calculation procedures, often accounting for 75% of the total effort. Estimates of V_{fb}, V_{fs}, and V_f are not changed if data are analysed in terms of plot means rather than individual tree means, but V_e cannot be determined. The term (V_e/nbs) is often so small that it is inconsequential in the estimation of family heritability. However, single tree heritability is slightly overestimated if V_e is omitted. Even more time can be saved by dealing solely with the means of families at different sites, *i.e.*, calculating V_{fs} and V_f only. Elimination of the V_{fb}/bs term ordinarily causes a slight overestimate of family heritability. Elimination of the V_{fb} term may cause a greater overestimate of the single tree heritability.

Step 5. Calculate the standard error of single tree heritability estimate as

$$SE(h^2) = \frac{\left(1 - h^2/4\right)\left[1 + (nbs - 1)h^2/4\right]}{\left[\left(nbs/2\right)(nbs - 1)(f - 1)\right]^{\frac{1}{2}}} \qquad (6.11)$$

$$= \frac{\left(1 - 0.5676/4\right)\left[1 + ((6)(3)(2) - 1)0.5676/4\right]}{\left[\left((6)(3)(2)/2\right)((6)(3)(2) - 11)(6 - 1)\right]^{\frac{1}{2}}}$$

$$= 0.0036$$

The standard error of family heritability is approximately given by

$$SE(h^2) \cong \frac{(1-t)(1+nbst)}{\left[(nbs)(f-1)/2\right]^{\frac{1}{2}}} \qquad (6.12)$$

$$\cong \frac{(1-0.1419)\left(1+(6)(3)(2)(0.1419)\right)}{\left[\left((6)(3)(2)\right)(6-1)/2\right]^{\frac{1}{2}}}$$

$$\cong 0.5525$$

where t is the intraclass correlation, which equals one-fourth of the single tree heritability.

The above formulae are correct if $V_e = V_{fb} = V_{fs}$. However, if one of these is much larger than the others, the term nbs should be reduced accordingly. If, for example, V_{fs} is much larger than V_{fb} or V_e, s might be substituted for nbs.

The above-calculated family heritability estimate is strictly applicable only if those families with the best overall performance in all plantations are selected. A breeder may select those families which are superior in one plantation only. In that case, family heritability is calculated as above except that V_{fs} is substituted for V_{fs}/s in the denominator.

If a breeder wishes to select on the basis of plot means, only family heritability is calculated as shown above, except that V_{fs} and V_{fb} are substituted for V_{fs}/s and V_{fb}/bs, respectively, in the denominator.

Step 6. To calculate genetic gain from a half-sib progeny test, use the formula for genetic gain from family selection.

Genetic gain = Selection differential x Family heritability (6.13)
where Selection differential = (Mean of selected families - Mean of all families)

To calculate expected gain from mass selection in such a progeny test, use the formula,

Expected mass selection gain = Selection differential x Single tree heritability (6.14)
where Selection differential = (Mean of selected trees - Mean of all trees)

6.1.2. Genotype-environment interaction

The phenotype of an individual is the resultant effect of its genotype and the environment in which it develops. Furthermore, the effects of genotype and environment may not be independent. A specific difference in environment may have a greater effect on some genotypes than on others, or there may be a change in the ranking of genotypes when measured in diverse environments. This interplay of genetic and non-genetic effects on the phenotype expression is called genotype-environment

interaction. The failure of a genotype to give the same response in different environments is a definite indication of genotype-environment interaction.

The environment of an individual is made up of all the things other than the genotype of the individual that affect its development. That is to say, environment is the sum total of all non-genetic factors external to the organism. Comstock and Moll (1963) distinguish two kinds of environments, micro and macro. Micro-environment is the environment of a single organism, as opposed to that of another, growing at the same time and in almost the same place. Specifically, micro-environmental differences are environmental fluctuations which occur even when individuals are apparently treated alike. On the other hand, environments that are potential or realized within a given area and period of time are referred to collectively as a macro-environment. A macro-environment can thus be conceived of as a collection of micro-environments that are potential therein. Different locations, climates and even different management practices are examples of macro-environmental differences. It is to be noted that the effect of micro-environment on an organism as well as its interactions with different genotypes is usually very small. Moreover, owing to the unpredictable and uncontrollable nature of micro-environment, its interactions with genotypes cannot be properly discerned. In other words, it is the macro-environmental deviation and its interaction with genotype that can be isolated and tested for significance.

One method of detecting genotype-environment interaction is by analysing data from a multi-location trial as done in Table 6.2 and testing the significance of the Family x Site interaction term. The computed F value is compared against table value of F for$(f\text{-}1)(s\text{-}1)$ and $s(f\text{-}1)(b\text{-}1)$ degrees of freedom (See Table 6.5).

If the interaction is nonsignificant or does not involve appreciable differences in rank among the best families or clones, they may be ignored, in which case selections should be based upon a genotype's average performance at all test sites. If the interactions are large and can be interpreted sufficiently to permit forecasts of where particular genotypes will excel or grow poorly, they cannot be ignored. To determine this, group the data from several plantations according to the plantations' site characteristics (*i.e.*, northern versus southern, dry versus moist, infertile versus fertile). Determine the amount of interaction within and between such groups. If a large portion of the interaction can be explained by the grouping, make separate selections for the sites typical of each plantation group. Then the correct statistical procedure is to make a separate analysis of variance and develop heritability estimate for each plantation group within which the interactions are too small or too uninterpretable to be of practical importance.

Table 6.5. Analysis of variance for a multi-plantation half-sib progeny test.

Sources of variation	Degrees of freedom	Sum of squares	Mean square	Computed F	Tabular F 5 %
Site	1	48900.46	48900.46		
Block-within-site	4	9258.13	2314.53		

Family	5	80533.37	16106.67		
Family x Site	5	35349.37	7069.87	$\frac{MSFS}{MSFB}$=3.97*	2.71
Family x Block-within-site	20	45183.87	2259.19		
Tree-within-plot	180	188799.67	1048.89		

* Significant at 5% level.

An alternative approach in this regard uses the regression technique in partitioning the genotype-environmental interaction component of variability into its linear and non-linear portions for assessing the stability of genotypes over a range of environments (Freeman and Perkins, 1971). Further discussion of this method is however not attempted here for want of space.

6.1.3. Seed orchard designs

A seed orchard is a plantation of genetically superior trees, isolated to reduce pollination from genetically inferior outside sources, and intensively managed to produce frequent, abundant, easily harvested seed crops. It is established by setting out clones (as grafts or cuttings) or seedling progeny of trees selected for desired characteristics. This section is concerned with certain specific planting designs used for raising seed orchards with emphasis on statistical aspects. Several other aspects of seed orchard planning related to type of planting material-clones or seedlings, number of clones or families, initial planting distances and related information can be found in books on tree breeding such as Wright, (1976) and Faulkner (1975).

In the case of clonal seed orchards, plants belonging to the same clone are called ramets. However, in this section, the term ‘clone’ or ‘ramet’, as applied in clonal seed orchards, are used for descriptive purposes. Similar designs can be used for seedling seed orchards, in which case the word ‘progeny’ should be substituted for ‘clone’ and ‘family-plot’ for ramet. Family plots can consist of a single tree or groups composed of several trees.

Completely randomized design (CRD) in which complete randomisation of all the available ramets of all clones between all the available planting positions on the site is the simplest of all designs to plan on paper. It can, however, pose practical management difficulties associated with planting, or, on-site grafting, and the relocation of individual ramets at a later stage and particularly when the orchard is large and contains many clones. If systematic thinning is to be practised by removing every second tree or every second row, the design can be further refined by making separate randomizations for the ramets which are to remain and for those to be removed in thinning. Quite frequently, certain restrictions are imposed on randomization, for example, that no two ramets of the same clone may be planted in adjacent positions within rows or columns, or where they will occur in adjacent diagonal positions; or, that at least two different ramets must separate ramets of the same clone. These restrictions are usually arranged by manipulating the positions of the ramets on the plan, thus making the design no

longer truly random, however, such deviations from randomness are seldom great. This strategy is adopted mainly to avoid the chances of inbreeding.

As an illustration, graphical layout of a completely randomized design for 10 clones with around 10 replications, planted with one-ring isolation is shown below.

4	7	4	8	5	10	7	6	4	7
8	3	9	1	2	1	3	5	3	5
6	1	5	3	10	5	10	9	7	10
8	4	2	1	9	7	6	3	5	8
5	7	3	6	2	3	5	2	10	2
1	10	4	7	10	6	8	4	1	5
9	7	6	3	5	2	7	3	6	2
1	5	2	10	1	3	10	5	4	9
8	10	4	7	5	7	8	2	1	6
7	2	8	6	1	4	6	7	10	4

Figure 6.1. Layout of a CRD for 10 clones with around 10 replications, with one-ring isolation around ramets of each clone.

As an extension of the above concepts, randomised complete block design (RCBD) or incomplete block designs like lattice designs discussed in Chapter 4 of this manual can be utilized in this connection for the benefits they offer in controlling the error component. However, the randomization within blocks is usually modified in order to satisfy restrictions on the proximity of ramets of the same clone. These designs are better suited for comparative clonal studies. Their main disadvantages are that RCBD shall not work well with large number of clones; lattice designs and other incomplete block designs are available only for certain fixed combinations of clone numbers and number of ramets per clone and they are unsuitable for systematic thinnings which would spoil the design.

La Bastide (1967) developed a computer programme which provides a design if feasible, for a set numbers of clones, ramets per clone, and ratio of rows to columns. There are two constraints; first, that there is a double ring of different clones to isolate each ramet of same clone (which are planted in staggered rows); second, that any combination of two adjacent clones should occur in any specific direction once only (See Figure 6.2). The design is called permutated neighbourhood design.

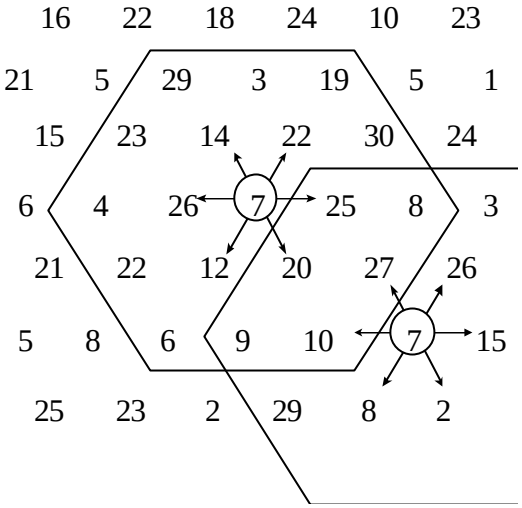


Figure 6.2. A fragment of a permutated neighbourhood design for 30 clones, with the restrictions on randomness employed by La Bastide (1967) in his computer design, *viz.*, (i) 2 rings of different clones isolate each ramet, and, (ii) any combination of two adjacent clones must not occur more than once in any specific direction.

Ideally, the design should be constructed for number of replications equal to one less than the number of clones, which would ensure that every clone has every other clone as neighbour once in each of the six possible directions. Thirty clones would therefore, require 29 ramets per clone or a total of 870 grafts although it may not be feasible to construct such large designs always. Even so, the small blocks which have been developed are, at the moment, the best designs available for ensuring, at least in theory, the maximum permutation of neighbourhood combinations and the minimum production of full-sibs in the orchard progeny. Chakravarty and Bagchi (1994) and Vanclay (1991) describe efficient computer programmes for construction of permutated neighbourhood seed orchard designs.

Seed orchards are usually established on the assumption that each clone and ramet, or, family-plot or seedling tree, in the orchard will: flower during the same period; will have the same cycle of periodic heavy flower production; be completely inter-fertile with all its neighbours and yield identical number of viable seed per plant; have the same degree of resistance to self-incompatibility; and will have a similar rate of growth and crown shape as all other plants. It is common experience that this is never the case and it is not likely to ever be so. The successful breeder will be the one who diligently observes and assiduously collects all essential information on clonal behaviour, compatibilities and combining-abilities, and translates this information into practical terms by employing it in the next and subsequent generations of seed orchards. Such designs will make maximum use of the available data.

6.2. Forest mensuration

6.2.1. Volume and biomass equations

In several areas of forestry research such as in silviculture, ecology or wood science, it becomes necessary to determine the volume or biomass of trees. Many times, it is the volume or biomass of a specified part of the tree that is required. Since the measurement of volume or biomass is destructive, one may resort to pre-established volume or biomass prediction equations to obtain an estimate of these characteristics. These equations are found to vary from species to species and for a given species, from stand to stand. Although the predictions may not be accurate in the case of individual trees, such equations are found to work well when applied repeatedly on several trees and the results aggregated, such as in the computation of stand volume. Whenever, an appropriate equation is not available, a prediction equation will have to be established newly. This will involve determination of actual volume or biomass of a sample set of trees and relating them to nondestructive measures like diameter at breast-height and height of trees through regression analysis.

(i) Measurement of tree volume and biomass

Determination of volume of any specified part of the tree such as stem or branch is usually achieved by cutting the tree part into logs and making measurements on the logs. For research purposes, it is usual to make the logs 3m in length except the top end log which may be up to 4.5m. But if the end section is more than 1.5m in length, it is left as a separate log. The diameter or girth of the logs is measured at the middle portion of the log, at either ends of the log or at the bottom, middle and tip potions of the logs depending on the resources available. The length of individual logs is also measured. The measurements may be made over bark or under bark after peeling the bark as required. The volume of individual logs may be calculated by using one of the formulae given in the following table depending on the measurements available.

Volume of the log	Remarks
$\frac{(b^2+t^2)l}{8\pi}$	Smalian's formula
$\left(\frac{m^2}{4\pi}\right)l$	Huber's formula
$\frac{\left(b^2+4m^2+t^2\right)l}{24\pi}$	Newton's formula

where b is the girth of the log at the basal portion
m is the girth at the middle of the log
t is the girth at the thin end of the log
l is the length of the log or height of the log

For illustrating the computation of volume of a tree using the above formulae, consider the data on bottom, middle, tip girth and length of different logs from a tree (Table 6.6).

Table 6.6. Bottom girth, middle girth, tip girth and length of logs of a teak tree.

Log number	Girth (cm)				Volume of log $(cm)^3$		
	Bottom (*b*)	Middle (*m*)	Tip (*t*)	Length (*l*)	Smalian's formula	Huber's formula	Newton's formula
1	129.00	99.00	89.00	570.00	556831.70	444386.25	481868.07
2	89.00	90.10	91.00	630.00	405970.57	406823.00	406538.86
3	64.00	60.00	54.90	68.00	19229.35	19472.73	19391.60
4	76.00	85.00	84.60	102.00	52467.48	58621.02	56569.84
5	84.90	80.10	76.20	111.00	57455.84	56650.45	56918.91
Total					1091954.94	985953.45	1021287.28

The volumes of individual logs are added to get a value for the volume of the tree or its part considered. In order to get the volume in m^3, the volume in $(cm)^3$ is to be divided by 1000,000.

Though volume is generally used in timber trade, weight is also used in the case of products like firewood or pulpwood. Weight is the standard measure in the case of many minor forest produce as well. For research purposes, biomass is getting increasingly more in use. Though use of weight as a measure appears to be easier than the use of volume, the measurement of weight is beset with problems like varying moisture content and bark, which render its use inconsistent. Hence biomass is usually expressed in terms of dry weight of component parts of trees such as stem, branches and leaves. Biomass of individual trees are determined destructively by felling the trees and separating the component parts like main stem, branches, twigs and leaves. The component parts are to be well defined as for instance, material below 10 cm girth over bark coming from main stem is included in the branch wood. The separated portions should be weighed immediately after felling. If oven-dry weights are needed, samples should be taken at this stage. At least three samples of about 1 kg should be taken from stem, branches and twigs from each tree. They should be weighed and then taken to the laboratory for oven-drying. The total dry weight of each component of the tree is then estimated by applying the ratio of fresh weight to dry weight observed in the sample to the corresponding total fresh weight of the component parts. For example,

$$\text{Total DW of bole} = \frac{\text{DW of samples from bole}}{\text{FW of samples from bole}}(\text{Total FW of bole}) \qquad (6.15)$$

where FW = Fresh weight
DW = Dry weight

For illustration, consider the data in Table 6.7.

Table 6.7. Fresh weight and dry weight of sample discs from the bole of a tree

Disc	Fresh weight (kg)	Dry weight (kg)
1	2.0	0.90
2	1.5	0.64
3	2.5	1.37
Total	6.0	2.91

$$\text{Total DW of bole} = \frac{\text{DW of samples from bole}}{\text{FW of samples from bole}}(\text{Total FW of bole})$$

$$\text{Total DW of bole of the tree} = \left(\frac{2.91}{6.00}\right)950$$
$$= 460.8 \text{ kg}$$

(ii) Estimation of allometric equations

The data collected from sample trees on their volume or biomass along with the dbh and height of sample trees are utilized to develop prediction equations through regression techniques. For biomass equations, sometimes diameter measured at a point lower than breast-height is used as regressor variable. Volume or biomass is taken as dependent variable and functions of dbh and height form the independent variables in the regression. Some of the standard forms of volume or biomass prediction equations in use are given below.

$$y = a + b\,D + c\,D^2 \quad (6.16)$$
$$\ln y = a + b\,D \quad (6.17)$$
$$\ln y = a + b \ln D \quad (6.18)$$
$$y^{0.5} = a + b\,D \quad (6.19)$$
$$y = a + b\,D^2H \quad (6.20)$$
$$\ln y = a + b\,D^2H \quad (6.21)$$
$$y^{0.5} = a + b\,D^2H \quad (6.22)$$
$$\ln y = a + b \ln D + c \ln H \quad (6.23)$$
$$y^{0.5} = a + b\,D + c\,H \quad (6.24)$$
$$y^{0.5} = a + b\,D^2 + c\,H + d\,D^2H \quad (6.25)$$

In all the above equations, y represents tree volume or biomass, D is the tree diameter measured at breast-height or at a lower point but measured uniformly on all the sample trees, H is the tree height, and a, b, c are regression coefficients, ln indicates natural logarithm.

Usually, several forms of equations are fitted to the data and the best fitting equation is selected based on measures like adjusted coefficient of determination or Furnival index. When the models to be compared do not have the same form of the dependent variable, Furnival index is invariably used.

$$\text{Adjusted } R^2 = 1 - \frac{n\text{-}1}{n\text{-}p}(1\text{-}R^2) \quad (6.26)$$

where R^2 is the coefficient of determination obtained as the ratio of regression sum of squares to the total sum of squares (see Section 3.7)
n is the number of observations on the dependent variable
p is the number of parameters in the model

Furnival index is computed as follows. The value of the square root of error mean square is obtained for each model under consideration through analysis of variance.

The geometric mean of the derivative of the dependent variable with respect to y is obtained for each model from the observations. Geometric mean of a set of n observations is defined by the nth root of the product of the observations. The Furnival index for each model is then obtained by multiplying the corresponding values of the square root of mean square error with the inverse of the geometric mean. For instance, the derivative of ln y is $(1/y)$ and the Furnival index in that case would be

$$\text{Furnival index} = \sqrt{MSE}\left(\frac{1}{\text{Geometric mean}\left(y^{-1}\right)}\right) \qquad (6.27)$$

The derivative of $y^{0.5}$ is $(1/2)(y^{-0.5})$ and corresponding changes will have to be made in Equation (6.27) when the dependent variable is $y^{0.5}$.

For example, consider the data on dry weight and diameter at breast-height of 15 acacia trees, given in Table 6.8.

Table 6.8. Dry weight and dbh of 15 acacia trees.

Tree no	Dry weight in tonne (y)	Dbh in metre (D)
1	0.48	0.38
2	0.79	0.47
3	0.71	0.44
4	1.86	0.62
5	1.19	0.54
6	0.51	0.38
7	1.04	0.50
8	0.62	0.43
9	0.83	0.48
10	1.19	0.48
11	1.03	0.52
12	0.61	0.40
13	0.68	0.44
14	0.20	0.26
15	0.66	0.44

Using the above data, two regression models, $y = a + b\,D + c\,D^2$ and $\ln y = a + b\,D$ were fitted using multiple regression analysis described in Montgomery and Peck (1982). Adjusted R^2 and Furnival index were calculated for both the models. The results are given in Tables 6.9 to Tables 6.12.

Table 6.9. Estimates of regression coefficients along with the standard error for the regression model, $y = a + b\,D + c\,D^2$.

Regression coefficient	Estimated regression coefficient	Standard error of estimated coefficient
a	0.5952	0.4810
b	-3.9307	2.0724

c	9.5316	2.4356

Table 6.10. ANOVA table for the regression analysis using the model, $y = a + b\,D + c\,D^2$.

Source	*df*	*SS*	*MS*	Computed *F*
Regression	2	2.0683	1.0341	105.6610
Residual	12	0.1174	0.0098	

$$R^2 = \frac{SSR}{SSTO} = \frac{2.0683}{2.1857} = 0.9463$$

$$\text{Adjusted } R^2 = 1 - \frac{15-1}{15-3}(1-0.9463)$$
$$= 0.9373$$

Here the derivative of y is 1. Hence,

$$\text{Furnival index} = \sqrt{MSE} = \sqrt{0.0098} = = 0.0989.$$

Table 6.11.Estimates of regression coefficients along with the standard error for the regression model $\ln y = a + b\,D$.

Regression coefficient	Estimated regression coefficient	Standard error of estimated coefficient
a	-3.0383	0.1670
b	6.0555	0.3639

Table 6.12. ANOVA table for the regression analysis using the model, $\ln y = a + b\,D$

Source	*df*	*SS*	*MS*	Computed *F*
Regression	1	3.5071	3.5071	276.9150
Residual	13	0.1646	0.0127	

$$R^2 = \frac{SSR}{SSTO} = \frac{3.5071}{3.5198} = 0.9552$$

$$\text{Adjusted } R^2 = 1 - \frac{15-1}{15-2}(1-0.9552)$$
$$= 0.9517$$

Here derivative of y is $1/y$. Hence, Furnival index calculated by Equation (6.27) is

$$\text{Furnival index} = \sqrt{0.0127}\left(\frac{1}{1.3514}\right) = 0.0834$$

Here, the geometric mean of $(1/y)$ will be the geometric mean of the reciprocals of the fifteen y values in Table 6.8.

In the example considered, the model ln $y = a + b\,D$ has a lower Furnival index and so is to be preferred over the other model $y = a + b\,D + c\,D^2$. Incidentally, the former model also has a larger adjusted R^2.

6.2.2. Growth and yield models for forest stands

Growth and yield prediction is an important aspect in forestry. 'Growth' refers to irreversible changes in the system during short phases of time. 'Yield' is growth integrated over a specified time interval and this gives the status of the system at specified time points. Prediction of growth or yield is important because many management decisions depend on it. For instance, consider the question, is it more profitable to grow acacia or teak in a place? The answer to this question depends, apart from the price, on the expected yield of the species concerned in that site. How frequently should a teak plantation be thinned ? The answer naturally depends on the growth rate expected of the plantation concerned. What would be the fate of teak if grown mixed with certain other species ? Such questions can be answered through the use of appropriate growth models.

For most of the modelling purposes, stand is considered as a unit of management. 'Stand' is taken as a group of trees associated with a site. Models try to capture the stand behaviour through algebraic equations. Some of the common measures of stand attributes are described here first before discussing the different stand models.

(i) Measurement of stand features

The most common measurements made on trees apart from a simple count are diameter at breast height or girth at breast-height and total height. Reference is made to standard text books on mensuration for the definition of these terms (Chaturvedi and Khanna, 1982). Here, a few stand attributes that are derivable from these basic measurements and some additional stand features are briefly mentioned.

Mean diameter : It is the diameter corresponding to the mean basal area of a group of trees or a stand, basal area of a tree being taken as the cross sectional area at the breast-height of the tree.

Stand basal area : The sum of the cross sectional area at breast-height of trees in the stand usually expressed m^2 on a unit area basis.

Mean height : It is the height corresponding to the mean diameter of a group of trees as read from a height-diameter curve applicable to the stand.

Top height : Top height is defined as the height corresponding to the mean diameter of 250 biggest diameters per hectare as read from height diameter curve.

Site index : Projected top height of a stand to a base age which is usually taken as the age at which culmination of height growth occurs.

Stand volume : The aggregated volume of trees in the stand usually expressed in m^3 on a unit area basis.

According to the degree of resolution of the input variables, stand models can be classified as (i) whole stand models (ii) diameter class models and (iii) individual tree models. Though a distinction is made as models for even-aged and uneven-aged stands, most of the models are applicable for both the cases. Generally, trees in a plantation are mostly of the same age and same species whereas trees in natural forests are of different age levels and of different species. The term even-aged is applied to crops consisting of trees of approximately the same age but differences up to 25% of the rotation age may be allowed in case where a crop is not harvested for 100 years or more. On the other hand, the term uneven-aged is applied to crops in which the individual stems vary widely in age, the range of difference being usually more than 20 years and in the case of long rotation crops, more than 25% of the rotation.

Whole stand models predict the different stand parameters directly from the concerned regressor variables. The usual parameters of interest are commercial volume/ha, crop diameter and crop height. The regressor variables are mostly age, stand density and site index. Since age and site index determine the top height, sometimes only the top height is considered *in lieu* of age and site index. The whole stand models can be further grouped according to whether or not stand density is used as an independent variable in these models. Traditional normal yield tables do not use density since the word 'normal' implies Nature's maximum density. Empirical yield tables assume Nature's average density. Variable-density models split by whether current or future volume is directly estimated by the growth functions or whether stand volume is aggregated from mathematically generated diameter classes. A second distinction is whether the model predicts growth directly or uses a two-stage process which first predicts future stand density and then uses this information to estimate future stand volume and subsequently growth by subtraction.

Diameter class models trace the changes in volume or other characteristics in each diameter class by calculating growth of the average tree in each class, and multiply this average by the inventoried number of stems in each class. The volumes are the aggregated over all classes to obtain stand characteristics.
Individual tree models are the most complex and individually model each tree on a sample tree list. Most individual tree models calculate a crown competition index for each tree and use it in determining whether the tree lives or dies and, if it lives, its growth in terms of diameter, height and crown size. A distinction between model types is based on how the crown competition index is calculated. If the calculation is based on the measured or mapped distance from each subject tree to all trees with in its competition zone, then it is called distance-dependent. If the crown competition index is based only on the subject tree characteristics and the aggregate stand characteristics, then it is a distance-independent model.

A few models found suitable for even-aged and uneven-aged stands are described separately in the following.

(ii) *Models for even-aged stands*

Sullivan and Clutter (1972) gave three basic equations which form a compatible set in the sense that the yield model can be obtained by summation of the predicted growth through appropriate growth periods. More precisely, the algebraic form of the yield model can be derived by mathematical integration of the growth model. The general form of the equations is

$$\text{Current yield} = V_1 = f(S, A_1, B_1) \quad (6.28)$$
$$\text{Future yield} = V_2 = f(S, A_2, B_2) \quad (6.29)$$
$$\text{Projected basal area} = B_2 = f(A_1, A_2, S, B_1) \quad (6.30)$$

where S = Site index
V_1 = Current stand volume
V_2 = Projected stand volume
B_1 = Current stand basal area
B_2 = Projected stand basal area
A_1 = Current stand age
A_2 = Projected stand age

Substituting Equation (6.30) for B_2 in Equation (6.29), we get an equation for future yield in terms of current stand variables and projected age,

$$V_2 = f(A_1, A_2, S, B_1) \quad (6.31)$$

A specific example is

$$\log V_2 = \beta_0 + \beta_1 S + \beta_2 A_2^{-1} + \beta_3\left(1 - A_1 A_2^{-1}\right) + \beta_4\left(\log B_1\right) A_1 A_2^{-1} \quad (6.32)$$

The parameters of Equation (6.32) can be estimated directly through multiple linear regression analysis (Montgomery and Peck,1982) with remeasured data from permanent sample plots, keeping V_2 as the dependent variable and A_1, A_2, S and B_1 as independent variables.
Letting $A_2 = A_1$, in Equation (6.32),

$$\log V = \beta_0 + \beta_1 S + \beta_2 A^{-1} + \beta_3 \log B \quad (6.33)$$

which is useful for predicting current volume.

To illustrate an application of the modelling approach, consider the equations reported by Brender and Clutter (1970) which was fitted to 119 remeasured piedmont loblolly pine stands near Macon, Georgia. The volume (cubic foot/acre) projection equation is

$$\log V_2 = 1.52918 + 0.002875S + 6.1585A_2^{-1} + 2.291143\left(1 - A_1 A_2^{-1}\right) + 0.93112\left(\log B_1\right) A_1 A_2^{-1} \quad (6.34)$$

Letting $A_2 = A_1$, this same equation predicts the current volume as,

$$\log V = 1.52918 + 0.002875S - 6.15851A^{-1} + 0.93112(\log B) \qquad (6.35)$$

To illustrate an application of the Brender-Clutter model, assume a stand growing on a site of site index of 80 feet, currently 25 years old with a basal area of 70 ft^2/acre. The owner wants an estimate of current volume and the volume expected after 10 more years of growth. Current volume is estimated using Equation (6.35),

$$\log V = 1.52918 + 0.002875(80) - 6.15851(1/25) + 0.93112(\log 70)$$
$$= 1.52918 + 0.23 - 0.24634 + 1.71801$$
$$= 3.23085$$

$V = 10^{3.23085} = 1{,}701\ ft^3$.

Volume after 10 years would be, as per Equation (6.34),

$$\log V_2 = 1.52918 + 0.002875(80) + 6.1585(1/25) + 2.291143(1 - 25/35) + 0.93112(\log 70)(25/35)$$
$$= 1.52918 + 0.23 - 0.24634 + 0.65461 - 1.22714$$
$$= 3.39459$$

$V_2 = 2{,}480\ ft^3$

(iii) Models for uneven-aged stands

Boungiorno and Michie (1980) present a matrix model in which the parameters represent (i) stochastic transition of trees between diameter classes and (ii) ingrowth of new trees which depends upon the condition of the stand. The model has the form

$$y_{1t+\vartheta} = \beta_0 + g_1(y_{1t} - h_{1t}) + g_2(y_{2t} - h_{2t}) + \ldots + g_n(y_{nt} - h_{nt}) \qquad (6.36)$$
$$y_{2t+\theta} = b_2(y_{1t} - h_{1t}) + a_2(y_{2t} - h_{2t})$$
$$\cdots$$
$$y_{nt+\theta} = b_n(y_{\{n-1\}t} - h_{\{n-1\}t}) + a_n(y_{nt} - h_{nt})$$

where $y_{it+\vartheta}$ gives the expected number of living trees in the ith size class at time t.

h_{it} gives the number of trees harvested from ith size classes during a time interval.

g_i, a_i, b_i are coefficients to be estimated.

Here the number of trees in the smallest size class is expressed as a function of the number of trees in all size classes and of the harvest within a particular time interval. With the same time reference, the numbers of trees in higher size classes are taken as functions of the numbers of trees in adjacent size classes. It is possible to estimate the parameters through regression analysis using data from permanent sample plots wherein status of the number of trees in different diameter classes in each time period with a specified interval is recorded along with the number of trees harvested between successive measurements.

For an over-simplified illustration, consider the following data collected at two successive instances with an interval of $\theta = 5$ years from a few permanent sample plots in natural forests. The data given in Table 6.13 show the number of trees in three diameter classes at the two measurement periods. Assume that no harvesting has taken place during the interval, implying h_{it}; $i = 1, 2, \ldots, n$ to be zero. In actual applications, more than three diameter classes may be identified and data from multiple measurements from a large number of plots will be required with records of number of trees removed from each diameter class between successive measurements.

Table 6.13. Data on number of trees/ha in three diameter classes at two successive measurements in natural forests.

Sample plot number	Number of trees /ha at Measurement - I			Number of trees/ha at Measurement - II		
	dbh class <10cm (y_{1t})	dbh class 10-60 cm (y_{2t})	dbh class >60 cm (y_{3t})	dbh class <10cm (y_{1t+q})	dbh class 10-60 cm (y_{2t+q})	dbh class >60 cm (y_{2t+q})
1	102	54	23	87	87	45
2	84	40	22	89	71	35
3	56	35	20	91	50	30
4	202	84	42	77	167	71
5	34	23	43	90	31	29
6	87	23	12	92	68	20
7	78	56	13	90	71	43
8	202	34	32	82	152	33
9	45	45	23	91	45	38
10	150	75	21	83	128	59

The equations to be estimated are,

$$y_{1t+\vartheta} = \beta_0 + g_1 y_{1t} + g_2 y_{2t} + g_3 y_{3t} \qquad (6.37)$$
$$y_{2t+\theta} = b_2 y_{1t} + a_2 y_{2t}$$
$$y_{3t+\theta} = b_3 y_{2t} + a_3 y_{3t}$$

Assembling the respective data from Table 6.13 and running the multiple linear regression routine (Montgomery and Peck,1982), the following estimates may be obtained.

$$y_{1t+\vartheta} = 99.8293 - 0.0526y_{1t} - 0.0738y_{2t} - 0.1476y_{3t} \quad (6.38)$$
$$y_{2t+\theta} = 0.7032y_{1t} + 0.2954y_{2t}$$
$$y_{3t+\theta} = 0.7016y_{2t} + 0.2938y_{3t}$$

Equations such as in model (6.38) have great importance in projecting the future stand conditions and devising optimum harvesting policies on the management unit as demonstrated by Boungiorno and Michie (1980). Growth models in general are put to use in forest management for comparing alternative management prescriptions. Using growth simulation models, it is possible to compare the simulated out-turn under the different prescriptions with respect to measures like internal rate of return (IRR) and arrive at optimum harvesting schedules. As growth and yield projections can be made under a variety of models, a choice will have to be made with respect to the best model to be used for such a purpose. Models differ with respect to data requirement or computational complexities. Apart from these, the biological validity and accuracy of prediction are of utmost importance in the choice of a model.

6.3. Forest ecology

6.3.1. Measurement of biodiversity

Biodiversity is the property of living systems of being distinct, that is different, unlike. Biological diversity or biodiversity is defined here as the property of groups or classes of living entities to be varied. Biodiversity manifests itself in two dimensions *viz*., variety and relative abundance of species (Magurran, 1988). The former is often measured in terms of species richness index which is,

$$\text{Species richness index} = \frac{S}{\sqrt{N}} \quad (6.39)$$

where S = Number of species in a collection
N = Number of individuals collected

As an illustration, suppose we encounter 400 species in a collection of 10000 individuals, the species richness index would be.

$$\text{Species richness index} = \frac{400}{\sqrt{10000}} = 4$$

The increase in the number of species in relation to the number of individuals or the area covered is represented by a species accumulation curve. The relation between number of species (S) and the area (A) covered is often represented mathematically by the equation, $S = \alpha A^{\beta}$, a graph of which is shown below for specific values of α and β ($\alpha = 100$, $\beta = 0.2$). Here, α and β are parameters to be estimated empirically using linear regression techniques with data on area covered and the corresponding number of species recorded.

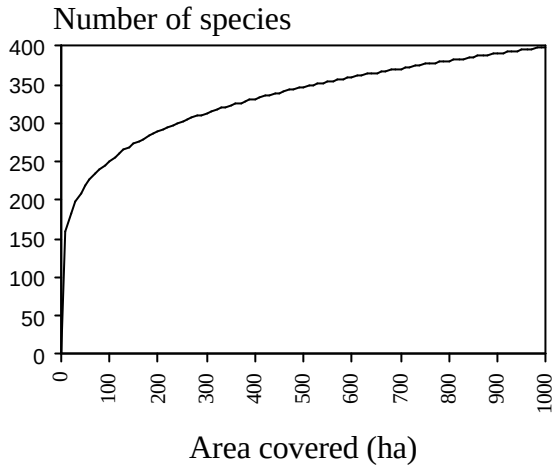


Figure 6.3. An example of species-area curve

Using the equation $S = 100A^{0.2}$, we shall be able to predict the number of possible species that we will get by covering a larger area within the region of sampling. In the above example, we are likely to get '458' species when the area of search is '2000 ha'.

In instances like the collection of insects through light traps, a species-individual curve will be more useful. In order to get an asymptotic curve we may have to use nonlinear equations of the form

$$S = \frac{\alpha N}{\beta + N} \tag{6.40}$$

wherein S tends to α as N tends to ∞. This means that α will be the limiting number of species in an infinitely large collection of individuals. The parameters α and β in this case will have to be estimated using nonlinear regression techniques (Draper and Smith, 1966). A graph of Equation (6.40) is shown below for $\alpha = 500$ and $\beta = 100$.

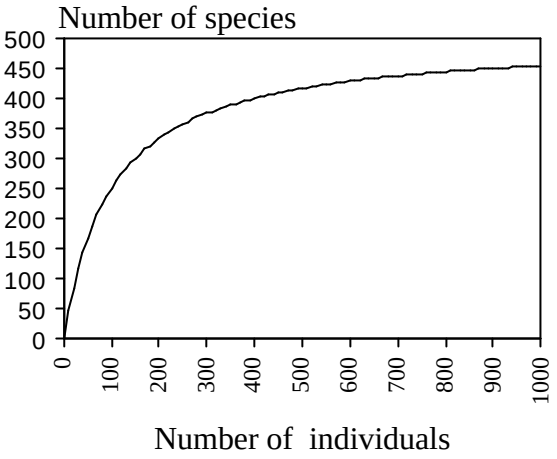


Figure 6.4. An example of species-individual curve

The relative abundance is usually measured in terms of diversity indices, a best known example of which is Shannon-Wiener index (*H*).

$$H = -\sum_{i=1}^{S} p_i \ln p_i \tag{6.41}$$

where p_i = Proportion of individuals found in the *i*th species
ln indicates natural logarithm

The values of Shannon-Wiener index obtained for different communities can be tested using Student's *t* test where *t* is defined as

$$t = \frac{\left|H_1 - H_2\right|}{\sqrt{Var(H_1) + Var(H_2)}} \tag{6.42}$$

which follows Student's *t* distribution with ν degrees of freedom where

$$\nu = \frac{(Var(H_1) + Var(H_2))^2}{(Var(H_1))^2 / N_1 + (Var(H_2))^2 / N_2} \tag{6.43}$$

$$Var(H) = \frac{\sum p_i (\ln p_i)^2 - (\sum p_i \ln p_i)^2}{N} + \frac{S-1}{2N^2} \tag{6.44}$$

The calculation of Shannon-Wiener index and testing the difference between the indices of two locations are illustrated below.

Table 6.14 shows the number of individuals belonging to different insect species obtained in collections using light traps at two locations in Kerala (Mathew *et al.*, 1998).

Table 6.14. Number of individuals belonging to different insect species obtained in collections using light traps at two locations

Species code	Number of individuals collected from Nelliampathy	Number of individuals collected from Parambikulum
1	91	84
2	67	60
3	33	40
4	22	26
5	27	24
6	23	20
7	12	16
8	14	13
9	11	12
10	10	7
11	9	5
12	9	5
13	5	9
14	1	4
15	4	6
16	2	2
17	2	4
18	1	4
19	2	5
20	4	1

Step 1. The first step when calculating the Shannon-Wiener index by hand is to draw up a table (Table 6.15) giving values of p_i and $p_i \ln p_i$. In cases where t test is also used, it is convenient to add a further column to the table giving values of $p_i (\ln p_i)^2$.

Step 2. The insect diversity in Nelliyampathy is $H_1 = 2.3716$ while the diversity in Parambikulam is $H_2 = 2.4484$. These values represent the sum of the $p_i \ln p_i$ column. The formula for the Shannon-Wiener index commences with a minus sign to cancel out the negative signs created by taking logarithms of proportions.

Step 3. The variance in diversity of the two locations may be estimated using Equation (6.44).

$$Var(H) = \frac{\sum p_i (\ln p_i)^2 - (\sum p_i \ln p_i)^2}{N} + \frac{S-1}{2N^2}$$

Thus, $Var(H_1)$ -Nelliyampathy = $\frac{6.6000 - 5.6244}{349} + \frac{19}{2(349)^2} = 0.0029$

$Var(H_2)$ -Parambikulam = $\frac{6.9120 - 5.9947}{347} + \frac{19}{2(347)^2} = 0.0027$

Table 6.15. Calculation of Shannon-Wiener index for two locations.

Species code	Nelliyampathy			Parambikulam		
	p_i	$p_i \ln p_i$	$p_i (\ln p_i)^2$	p_i	$p_i \ln p_i$	$p_i (\ln p_i)^2$
1	0.2607	-0.3505	0.4712	0.2421	-0.3434	0.4871
2	0.1920	-0.3168	0.5228	0.1729	-0.3034	0.5325
3	0.0946	-0.2231	0.5262	0.1153	-0.2491	0.5381
4	0.0630	-0.1742	0.4815	0.0749	-0.1941	0.5030
5	0.0774	-0.1980	0.5067	0.0692	-0.1848	0.4936
6	0.0659	-0.1792	0.4873	0.0576	-0.1644	0.4692
7	0.0344	-0.1159	0.3906	0.0461	-0.1418	0.4363
8	0.0401	-0.1290	0.4149	0.0375	-0.1231	0.4042
9	0.0315	-0.1090	0.3768	0.0346	-0.1164	0.3916
10	0.0286	-0.1016	0.3609	0.0202	-0.0788	0.3075
11	0.0258	-0.0944	0.3453	0.0144	-0.0611	0.2591
12	0.0258	-0.0944	0.3453	0.0144	-0.0611	0.2591
13	0.0143	-0.0607	0.2577	0.0259	-0.0946	0.3456
14	0.0029	-0.0169	0.0990	0.0115	-0.0514	0.2295
15	0.0115	-0.0514	0.2297	0.0173	-0.0702	0.2848
16	0.0057	-0.0294	0.1518	0.0058	-0.0299	0.154
17	0.0057	-0.0294	0.1518	0.0115	-0.0514	0.2295
18	0.0029	-0.0169	0.099	0.0115	-0.0514	0.2295
19	0.0057	-0.0294	0.1518	0.0144	-0.0611	0.2591
20	0.0115	-0.0514	0.2297	0.0029	-0.0169	0.0987
Total	1	-2.3716	6.6000	1	-2.4484	6.9120

Step 4. The *t* test allows the diversity of the two locations to be compared. The appropriate formulae are given in Equations (6.42) and (6.43).

$$t = \frac{|H_1 - H_2|}{\sqrt{Var(H_1) + Var(H_2)}}$$

$$\nu = \frac{(Var(H_1) + Var(H_2))^2}{(Var(H_1))^2 / N_1 + (Var(H_2)^2 / N_2}$$

In this example, $t = \frac{|2.3716 - 2.4484|}{\sqrt{0.0029 + 0.0027}} = 1.0263$

The corresponding degrees of freedom are calculated as

$$n = \frac{(0.0029 + 0.0027)^2}{(0.0029)^2 / 349 + (0.0027)^2 / 347} = 695.25$$

The table value of *t* corresponding to 695 degrees of freedom (Appendix 2) shows that the difference between diversity indices of two locations is nonsignificant.

Conventionally, random sampling patterns are employed in studies on biodiversity. One related question is, what is the sample size required to estimate any particular diversity index. Simulation exercises based on realistic structure of species abundances revealed that observing 1000 randomly selected individuals is adequate to estimate Shannon-Wiener index. Estimation of species richness may need an effort level of about 6000 (Parangpe and Gore, 1997).

6.3.2. Species abundance relation

A complete description of the relative abundance of different species in a community can be obtained through a species abundance model. The empirical distribution of species abundance is obtained by plotting the number of species against the number of individuals. Later, the observed distribution is approximated by a theoretical distribution. One of the theoretical models used in this connection especially with partially disturbed populations is the log series. The log series takes the form

$$\alpha x, \frac{\alpha x^2}{2}, \frac{\alpha x^3}{3}, \ldots, \frac{\alpha x^n}{n} \qquad (6.45)$$

αx being the number of species with one individual, $\alpha x^2/2$ the number of species with two individuals, etc. The total number of species (*S*) in the population is obtained by adding up all the terms in the series which will work out to $S = \alpha [- \ln (1-x)]$.

To fit the series, it is necessary to calculate how many species are expected to have one individual, two individuals and so on. These expected abundances are then put into the same abundance classes used for the observed distribution and a goodness of fit test is used to compare the two distributions. The total number of species in the observed and expected distributions is of course identical.

The calculations involved are illustrated with the following example. Mathew *et al.* (1998) studied the impact of forest disturbance on insect species diversity at four

locations in the Kerala part of Western Ghats. As part of their study, they assembled a list giving the abundance of 372 species at Nelliyampathy. This list is not reproduced here for want of space. However, this data set used here to illustrate the calculations involved in fitting a log series model.

Step 1. Put the observed abundances into abundance classes. In this case, classes in $\log_2$ (that is octaves or doublings of species abundances) are chosen. Adding 0.5 to the upper boundary of each class makes it straightforward to unambiguously assign observed species abundances to each class. Thus, in the table below (Table 6.16), there are 158 species with an abundance of one or two individuals, 55 species with an abundance of three or four individuals, and so on.

Table 6.16. Number of species obtained in different abundance classes.

Class	Upper boundary	Number of species observed
1	2.5	158
2	4.5	55
3	8.5	76
4	16.5	49
5	32.5	20
6	64.5	9
7	128.5	4
8	∞	1
Total number of species (*S*)	-	372

Step 2. The two parameters needed to fit the series are x and α. The value of x is estimated by iterating the following term.

$$\frac{S}{N} = [(1-x)/x][-\ln(1-x)] \qquad (6.46)$$

where S = Total number of species (372)
N = Total number of individuals (2804).

The value of x is usually greater than 0.9 and always <1.0. A few calculations on a hand calculator will quickly produce the correct value of x by trying different values of x in the expression $[(1-x)/x][-\ln(1-x)]$ and examining if it attains the value of $S/N = 0.13267$.

x	$[(1-x)/x][-\ln(1-x)]$
0.97000	0.10845
0.96000	0.13412
0.96100	0.13166
0.96050	0.13289
0.96059	0.13267

The correct value of x is therefore 0.96059. Once x has been obtained, it is simple to calculate α using the equation,

$$\alpha = \frac{N(1-x)}{x} = \frac{2804(1-0.96059)}{0.96059} = 115.0393 \qquad (6.47)$$

Step 3. When α and x have been obtained, the number of species expected to have 1, 2, 3, . . ., n individuals can be calculated. This is illustrated below for the first four abundance classes corresponding to the cumulative sums.

Table 6.17. Calculation involved in finding out the expected number of species in a log series model.

Number of Individuals	Series term	Number of expected species	Cumulative sum
1	αx	110.5	
2	$\alpha x^2/2$	53.1	163.6
3	$\alpha x^3/3$	33.9	
4	$\alpha x^4/4$	24.5	58.5
5	$\alpha x^5/5$	18.8	
6	$\alpha x^6/6$	15.1	
7	$\alpha x^7/7$	12.4	
8	$\alpha x^8/8$	10.4	56.7
9	$\alpha x^9/9$	8.9	
10	$\alpha x^{10}/10$	7.7	
11	$\alpha x^{11}/11$	6.7	
12	$\alpha x^{12}/12$	6.0	
13	$\alpha x^{13}/13$	5.2	
14	$\alpha x^{14}/14$	4.7	
15	$\alpha x^{15}/15$	4.2	
16	$\alpha x^{16}/16$	3.8	47.1

Step 4. The next stage is to compile a table giving the number of expected and observed species in each abundance class and compare the two distributions using a goodness of fit test. Chi-square test is one commonly used test. For each class, calculate χ^2 as shown.

χ^2 = (Observed frequency - Expected frequency)^2/ Expected frequency (6.48)

For example, in class 1, χ^2 = (158-163.5809)2 /163.5809 =0.1904. Finally sum this column to obtain the overall goodness of fit, $\sum \chi^2$. Check the obtained value in chi-square tables (Appendix 4) using (Number of classes-1) degrees of freedom. In this case, $\sum \chi^2 = 12.0624$, with 6 degrees of freedom. The value of χ^2 for P=0.05 is 12.592. We can therefore conclude that there is no significant

difference between the observed and expected distributions, *i.e*., the log series model fits well for the data.

If χ^2 is calculated when the number of expected species is small (<1.0) the resultant value of χ^2 can be extremely large. In such cases, it is best to combine the observed number of species in two or more adjacent classes and compare this with the combined number of expected species in the same two classes. The degrees of freedom should be reduced accordingly. In the above example, since the expected frequency of class 8 was less than 1, the observed and expected frequencies of class 8 were combined with those of class 7 while testing for goodness of fit.

Table 6.18. Test of goodness of fit of log series model.

Class	Upper boundary	Observed	Expected	(Observed - Expected)2 / Expected
1	2.5	158	163.5809	0.1904
2	4.5	55	58.4762	0.2066
3	8.5	76	56.7084	6.5628
4	16.5	49	47.1353	0.0738
5	32.5	20	30.6883	3.7226
6	64.5	9	11.8825	0.6992
7	128.5	5	3.5351	0.6070
Total		372	372.0067	12.0624

6.3.3. Study of spatial patterns

Spatial distribution of plants and animals is an important characteristic of ecological communities. This is usually one of the first observations that is made while studying any community and is one of the most fundamental properties of any group of living organisms. Once a pattern has been identified, the ecologist may propose and test hypotheses that explain the underlying causal factors. Hence, the ultimate objective of detecting spatial patterns is to generate hypothesis concerning the structure of ecological communities. In this section, the use of statistical distributions and a few indices of dispersion for detecting and measuring spatial pattern of species in communities are described.

Three basic types of patterns are recognised in communities *viz*., random, clumped and uniform (See Figure 6.5).The following causal mechanisms are often used to explain observed patterns in ecological communities. Random patterns in a population of organisms imply environmental homogeneity and/or non-selective behavioural patterns. On the other hand, non-random patterns (clumped and uniform) imply that some constraints on the population exist. Clumping suggests that individuals are aggregated in more favourable parts of the habitat; this may be due to gregarious behaviour, environmental heterogeneity, reproductive mode, and so on. Uniform dispersions result from negative interactions between individuals, such as competition for food or space. One has to note that detecting a pattern and explaining its possible causes are separate

problems. Furthermore, it should be kept in mind that nature is multifactorial; many interacting processes (biotic and abiotic) may contribute to the existence of patterns.

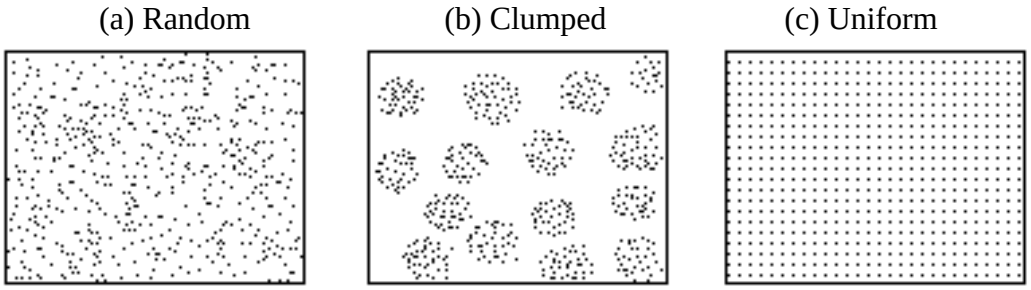


Figure 6.5. Three basic patterns of spatial distribution.

Hutchinson was one of the first ecologists to consider the importance of spatial patterns in communities and identify various causal factors that may lead to patterning of organisms like (i) vectorial factors resulting from the action of external environmental forces (*e.g.*,wind, water currents, and light intensity) (ii) reproductive factors attributable to the reproductive mode of the organism (*e.g*., cloning and progeny regeneration) (iii) social factors due to innate behaviours (*e.g*., territorial behaviour); (iv) coactive factors resulting from intra-specific interactions (*e.g*. competition); and (v) stochastic factors resulting from random variation in any of the preceding factors. Thus, processes contributing to spatial patterns may be considered as either intrinsic to the species (*e.g*., reproductive, social and coactive) or extrinsic (*e.g*., vectorial). Further discussions of the causes of pattern are given in Ludwig and Reynolds (1988).

If individuals of a species are spatially dispersed over discrete sampling units *e.g*., scale insects on plant leaves, and if at some point in time a sample is taken of the number of individuals per sampling unit, then it is possible to summarize these data in terms of a frequency distribution. This frequency distribution consists of the number of sampling units with 0 individual, 1 individual, 2 individuals, and so on. This constitutes the basic data set we use in the pattern detection methods described subsequently. Note that the species are assumed to occur in discrete sites or natural sampling units such as leaves, fruits, trees. Generally, it is observed that the relationships between the mean and variance of the number of individuals per sampling unit is influenced by the underlying pattern of dispersal of the population. For instance, mean and variance are nearly equal for random patterns, variance is larger than mean for clumped patterns and variance is less than mean for uniform patterns. There are certain statistical frequency distributions that, because of their variance-to-mean properties, have been used as models of these types of ecological patterns. These are (i) Poisson distribution for random patterns (ii) Negative binomial distribution for clumped patterns and (iii) Positive binomial for uniform patterns. While these three statistical models have commonly been used in studies of spatial pattern, it should be recognized that other statistical distributions might also be equally appropriate.

The initial step in pattern detection in community ecology often involves testing the hypothesis that the distribution of the number of individuals per sampling unit is

random. Poisson distribution has already been described in Section 2.4.2. If the hypothesis of random pattern is rejected, then the distribution may be in the direction of clumped (usually) or uniform (rarely). If the direction is toward a clumped dispersion, agreement with the negative binomial may be tested and certain indices of dispersion, which are based on the ratio of the variance to mean, may be used to measure the degree of clumping. Because of the relative rarity of uniform patterns in ecological communities, and also because the binomial distribution has been described earlier in Section 2.4.1, this case is not considered here.

Before proceeding, we wish to make some cautionary points. First of all, failure to reject a hypothesis of randomness means only that we have failed to detect non-randomness using the specified data set at hand. Second, we should propose only reasonable hypothesis in the sense that a hypothesis should be tenable and based on a mixture of common sense and biological knowledge. This second point has important ramifications with regard to the first. It is not uncommon for a theoretical statistical distribution (*e.g*., the Poisson series) to resemble an observed frequency distribution (*i.e*., there is a statistical agreement between the two) even though the assumptions underlying this theoretical model are not satisfied by the data set. Consequently, we may accept a null hypothesis that has, in fact, no biological justification. Third, we should not base our conclusions on significance tests alone. All available sources of information (ecological and statistical) should be used in concert. For example, failure to reject a null hypothesis that is based on a small sample size should be considered only as a weak confirmation of the null hypothesis. Lastly, it has to be remembered that the detection of spatial pattern and explaining its possible casual factors are separate problems.

The use of negative binomial distribution in testing for clumped patterns is described here. Negative binomial model is probably the most commonly used probability distribution for clumped, or what often are referred to as contagious or aggregated populations. When two of the conditions associated with the use of Poisson model are not satisfied, that is, condition 1 (each natural sampling unit has an equal probability of hosting an individual) and condition 2 (the occurrence of an individual in a sampling unit does not influence its occupancy by another), it usually leads to a high variance-to-mean ratio of the number of individuals per sampling unit. As previously shown, this suggests that a clumped pattern may exist.

The negative binomial has two parameters, μ, the mean number of individuals per sampling unit and k, a parameter related to the degree of clumping. The steps in testing the agreement of an observed frequency distribution with the negative binomial are outlined below.

Step 1. State the hypothesis ; The hypothesis to be tested is that the number of individuals per sampling unit follows a negative binomial distribution, and, hence, a nonrandom or clumped pattern exists. Failing to reject this hypothesis, the ecologist may have a good empirical model to describe a set of observed frequency data although this does not explain what underlying causal factors

might be responsible for the pattern. In other words, we should not attempt to infer causality solely based on our pattern detection approaches.

Step 2. The number of individuals per sampling unit is summarized as a frequency distribution, that is, the number of sampling units with 0, 1, 2, …, *r* individuals.

Step 3. Compute the negative binomial probabilities, *P(x)*. The probability of finding *x* individuals in a sampling unit, that is, *P(x)*, where x is 0, 1, 2, …, *r* individuals, is given by

$$P(x) = \left[\frac{\mu}{(\mu + k)}\right]^{x}\left[\frac{(k+x-1)!}{x!(k-1)!}\right]\left[1+\frac{\mu}{k}\right]^{-k} \qquad (6.49)$$

The parameter μ is estimated from the sample mean ($\bar{x}$). Parameter *k* is a measure of the degree of clumping and tends toward zero at maximum clumping. An estimate for *k* is obtained using the following iterative equation :

$$log_{10}\left(\frac{N}{N_0}\right) = \hat{k}\, log_{10}\left[1+\left(\frac{\bar{x}}{\hat{k}}\right)\right] \qquad (6.50)$$

where *N* is total number of sampling units in the sample and N_0 is the number of sampling units with 0 individuals. First, an initial estimate of $\hat{k}$ is substituted into the right-hand side (RHS). If the RHS is lower than the LHS, a higher value of $\hat{k}$ is then tried, and, again, the two sides are compared. This process is continued in an iterative fashion (appropriately selecting higher or lower values of $\hat{k}$) until a value of $\hat{k}$ is found such that the RHS converges to the same value as the LHS. A good initial estimate of $\hat{k}$ for the first iteration is obtained from,

$$\hat{k} = \frac{\bar{x}}{s^2 - \bar{x}} \qquad (6.51)$$

where s^2 is the sample estimate of variance.

When the mean is small (less than 4), Equation (6.50) is an efficient way to estimate $\hat{k}$. On the other hand, if the mean is large (greater than 4), this iterative method is efficient only if there is extensive clumping in the population. Thus, if both the population mean ($\bar{x}$) and the value of $\hat{k}$ (the clumping parameter, as computed from equation (6.51), are greater than 4, equation (6.51) is actually preferred over equation (6.50) for estimating $\hat{k}$.

Once the two statistics, $\bar{x}$ and $\hat{k}$, are obtained, the probabilities of finding x individuals in a sampling unit, that is, *P(x)*, where *x* = 0, 1, 2, …, *r* individuals, are computed from equation (6.49) as

$$P(0) = \left[\frac{\bar{x}}{\bar{x}+\hat{k}}\right]^0 \left\{\frac{(\hat{k}+0-1)!}{0!(k-1)!}\right\}\left[1+\frac{\bar{x}}{\hat{k}}\right]^{-k}$$

$$= \left[1+\left(\frac{\bar{x}}{\hat{k}}\right)\right]^{-k}$$

$$P(1) = \left[\frac{\bar{x}}{\bar{x}+\hat{k}}\right]^1 \left\{\frac{(\hat{k}+1-1)!}{1!(\hat{k}-1)!}\right\}\left[1+\left(\frac{\bar{x}}{\hat{k}}\right)\right]^{-k}$$

$$= \left[\frac{\bar{x}}{\bar{x}+\hat{k}}\right]\left(\frac{\hat{k}}{1}\right)P(0)$$

$$P(2) = \left[\frac{\bar{x}}{\bar{x}+\hat{k}}\right]^2 \left\{\frac{(\hat{k}+2-1)!}{2!(\hat{k}-1)!}\right\}\left[1+\left(\frac{\bar{x}}{\hat{k}}\right)\right]^{-k}$$

$$= \left[\frac{\bar{x}}{\bar{x}+\hat{k}}\right]\left(\frac{\hat{k}+1}{2}\right)P(1)$$

…
…
…

$$P(r) = \left[\frac{\bar{x}}{\bar{x}+\hat{k}}\right]^r \left\{\frac{(\hat{k}+r-1)!}{r!(\hat{k}-1)!}\right\}\left[1+\left(\frac{\bar{x}}{\hat{k}}\right)\right]^{-k}$$

$$= \left[\frac{\bar{x}}{\bar{x}+\hat{k}}\right]\left(\frac{\hat{k}+r-1}{r}\right)P(r-1)$$

Step 4. Obtain the expected negative binomial frequencies. The expected number of sampling units containing x individuals is obtained by multiplying each of the negative binomial probabilities by N, the total number of sampling units in the sample. The number of frequency classes, q, is also determined as described for the Poisson model.

Step 5.Perform a goodness of fit test. The chi-square test for goodness of fit is to be performed as described in Section 3.5.

An example of fitting negative binomial distribution is given in the following. Carpenter bee larvae are common in the inflorescence stalks of soap-tree yucca plants in southern New Mexico. An insect ecologist interested in the spatial patterns of these bees collected a random sample of bee larvae in 180 yucca stalks. These data, summarized in a frequency table, are

x	0	1	2	3	4	5	6	7	8	9	10
f_x	114	25	15	10	6	5	2	1	1	0	1

where x is the number of bee larvae per stalk and f_x is the frequency of yucca stalks having $x = 0, 1, 2, \ldots, r$ larvae. In this example, $r = 10$. The total number of sampling units is

$$N = \sum_{x=0}^{10}(f_x)$$
$$= 114 + 25 + \ldots..+ 0 + 1 = 180$$

and the total number of individuals is

$$n = \sum_{x=0}^{10}(xf_x) = (0)(114) + (1)(25) + (9)(0) + (10)(1) = 171$$

The arithmetic mean of the sample is

$$\bar{x} = \frac{n}{N} = \frac{171}{180}$$
$$= 0.95$$

and the variance is

$$s^2 = \frac{\left(\sum_{x=0}^{10}\left(xf_x\right)^2 - \bar{x}n\right)}{(n-1)}$$
$$= \frac{[681-(0.95)(171)]}{179}$$
$$= 2.897$$

Step 1. Hypothesis: The null hypothesis is that the carpenter bee larvae exhibit a clumped pattern in yucca inflorescence stalks and, hence, agreement (of the number of individuals per sampling unit) with a negative binomial is tested. Since the variance is greater than the mean, a clumped pattern is reasonably suspected.

Step 2. Frequency distribution, f_x : The observed frequency distribution, along with its mean and variance, are given previously.

Step 3. Negative binomial probabilities, $P(x)$: An estimate of $\hat{k}$ obtained using Equation.(6.51) with $\bar{x} = 0.95$ and $s^2 = 2.897$ is

$$\hat{k} = \frac{(0.95)^2}{(2.897 - 0.95)} = 0.4635$$

Since both $\hat{k}$ and $\bar{x}$ are less than 1, Equation (6.50) should be used to estimate $\hat{k}$. Substituting the values N =180 and N_0 =114 into the left-hand side (LHS) of Equation (6.50) gives the value of 0.1984. Now, substitution of $\hat{k}$ = 0.4635 into the right-hand side (RHS) of Equation (6.50) gives the following :

$$\textit{Iteration } 1: \hat{k}\log_{10}\left(1+\frac{\bar{x}}{\hat{k}}\right) = 0.4635\log_{10}\left(1+\frac{0.95}{0.4635}\right)$$
$$= 0.2245$$

Since the RHS is larger than 0.1984, a lower value than 0.4635 for $\hat{k}$ is now substituted into Equation (6.50). Selecting $\hat{k}$ = 0.30 gives,

$$\textit{Iteration } 2: \hat{k}\log_{10}\left(1+\frac{\bar{x}}{\hat{k}}\right) = 0.30\log_{10}\left(1+\frac{0.95}{0.30}\right)$$
$$= 0.1859$$

This is close to the value 0.1984 (but now lower), so for the next iteration, a slightly higher value of $\hat{k}$ is chosen. Using $\hat{k}$=0.34 gives

$$\textit{Iteration } 3: \hat{k}\log_{10}\left(1+\frac{\bar{x}}{\hat{k}}\right) = 0.34\log_{10}\left(1+\frac{0.95}{0.34}\right) = 0.1969$$

Again, for the next iteration, we try another value of $\hat{k}$ that is slightly higher. For $\hat{k}$=0.3457,

$$\textit{Iteration } 4: \hat{k}\log_{10}\left(1+\frac{\bar{x}}{\hat{k}}\right) = 0.3457\log_{10}\left(1+\frac{95}{0.3457}\right) = 0.1984$$

This is numerically identical to the LHS of Equation (6.50) and, for this example, the best estimate of $\hat{k}$ is 0.3457. Next, using equation (6.49), the individual and cumulative probabilities for finding 0, 1, 2, and 3 larvae per stalk [for $\bar{x}$=0.95 and $\hat{k}$=0.3457, where $\frac{\bar{x}}{(\bar{x}+\hat{k})} = 0.7332$] are given in Table 6.18.

The cumulative probabilities after finding the probability of 4 individuals in a sampling unit is 94.6%. Therefore, the remaining probabilities, $P(5)$ through $P(10)$ will contribute 5.4%, that is,

$P(5^+)$ = 1.0 - 0.946 = 0.054.

Table 6.18. Calculation for $P(x)$, the negative binomial probabilities, for x individuals (bees) per sampling unit (yucca stalks)

Probability		Cumulative probability
$P(0) = \left[1 + \left(\frac{0.95}{0.3457}\right)\right]^{-0.3457}$	=0.6333	0.6333
$P(1) = [0.7332]\left(\frac{0.3457}{1}\right)P(0) = (0.2535)(0.6333)$	=0.1605	0.7938
$P(2) = [0.7332]\left(\frac{1.3457}{2}\right)P(1) = (0.4933)(0.1605)$	=0.0792	0.8730
$P(3) = [0.7332]\left(\frac{2.3457}{3}\right)P(2) = (0.5733)(0.0792)$	=0.0454	0.9184
$P(4) = [0.7332]\left(\frac{3.3457}{4}\right)P(3) = (0.6133)(0.0454)$	=0.0278	0.9462
$P(5^+) = 1.00 - 0.9462$	=0.0538	1.0000

Step 4. Expected frequencies, E_x : The probabilities are multiplied by the total number of sampling units to obtain the expected frequencies (Table 6.19)

Table 6.19. Calculations for expected frequencies of sampling units containing different number of bees

Expected frequency				Cumulative frequency
E_0	$=(N)P(0)$	=(180)(0.633)	=114.00	114.00
E_1	$=(N)P(1)$	=(180)(0.161)	= 28.90	142.90
E_2	$=(N)P(2)$	=(180)(0.079)	= 14.25	157.20
E_3	$=(N)P(3)$	=(180)(0.045)	= 8.17	165.30
E_4	$=(N)P(4)$	=(180)(0.028)	= 5.00	170.30
E_{5+}	$=(N)P(5^+)$	=(180)(0.054)	= 9.68	180.00

Step 5.Goodness of fit : The chi-square test statistic (χ^2) is computed as,

$$\chi^2 = \left[\frac{(114 - 114.0)^2}{114.0}\right] + \ldots + \left[\frac{(10 - 9.67)^2}{9.67}\right]$$

$$= 0.00 + \ldots + 0.01 = 1.18$$

This value of the test statistic is compared to a table of critical values of the chi-square distribution with (Number of classes - 3) = 3 degrees of freedom. The critical value at the 5% probability level is 7.82 (Appendix 4), and since the probability of obtaining a value of χ^2 equal to 1.18 is well below this, we do not reject the null hypothesis. The negative binomial model appears to be a good fit to the observed data, but we would want further confirmation (*e.g.*, an independent set of data) before making definitive statements that the pattern of the carpenter been larvae is in fact, clumped. Note that when the minimal expected values are allowed to be as low as 1.0 and 3.0 in this example, the χ^2 values are 2.6 and 2.5, respectively - still well below the critical value.

Table 6.20. Calculations for χ^2 test statistic

Number of bee larvae per stalk (x)	Observed frequency f_x	Expected frequency E_x	$\frac{(f_x - E_x)^2}{E_x}$
0	114	114.0	0.00
1	25	28.9	0.53
2	15	14.3	0.04
3	10	8.2	0.41
4	6	5.0	0.19
5	10	9.7	0.01
Total	180	180.0	$\chi^2 = 1.18$

Other than using statistical distributions for detecting spatial patterns, certain easily computed indices can also be used for the purpose, like the index of dispersion or Green's index, when the sampling units are discrete.

(i) Index of dispersion : The variance-to-mean ratio or index of dispersion (*ID*) is

$$ID = \frac{s^2}{\bar{x}} \tag{6.52}$$

where $\bar{x}$ and s^2 are the sample mean and variance respectively. The variance-to-mean ratio (*ID*) is useful for assessing the agreement of a set of data to the Poisson series. However, in terms of measuring the degree of clumping, *ID* is not very useful. When the population is clumped, *ID* is strongly influenced by the number of individuals in the sample and, therefore, *ID* is useful as a comparative index of clumping only if n is the same in each sample. A modified version of *ID* that is independent of n is Green's index (*GI*) which is computed as,

$$GI = \frac{\left(\frac{s^2}{\bar{x}} - 1\right) - 1}{n - 1} \tag{6.53}$$

GI varies between 0 (for random) and 1 (for maximum clumping). Thus, Green's index can be used to compare samples that vary in the total number of individuals, their sample means, and the number of sampling in the sample. Consequently, of the numerous variants of *ID* that have been proposed to measure the degree of clumping, *GI* seems most recommendable. The values of *GI* for the scale insect population may be obtained as

$$GI = \frac{(3.05 - 1)}{(171 - 1)} = 0.012$$

Since the maximum value of GI is 1.0 (if all 171 individuals had occurred in a single yucca stalk), this value represents a relatively low degree of clumping.

6.3.4. Dynamics of ecosystems

It is well known that forests, considered as ecosystems exhibit significant changes over time. An understanding of these dynamic processes is important both from scientific and management perspectives. One component of these, the prediction of growth and yield of forests has received the greatest attention in the past. However, there are several other equally important aspects concerned with dynamics of forests like long term effects of environmental pollution, successional changes in forests, dynamics, stability and resilience of both natural and artificial ecosystems etc. These different application purposes require widely different modelling approaches. Complexity of these models does not even permit an overview of these models here and what has been attempted is just a simplified description of some of the potential models in this context.

Any dynamic process is shaped by the characteristic time-scale of its components. In forests, these range from minutes (stomatal processes) to hours (diurnal cycle, soil water dynamics), to days (nutrient dynamics, phenology), to months (seasonal cycle, increment), to years (tree growth and senescence), to decades (forest succession) and to centuries (forest response to climatic change). The model purpose determines which of these time-scales will be emphasized. This usually requires an aggregated description of processes having different time-scales, but the level of aggregation will depend on the degree of behavioural validity required.

The traditional method of gathering data to study forest dynamics at a macro level is to lay out permanent sample plots and make periodical observations. More recently, remote sensing through satellites and other means has offered greater scope for gathering accurate historical data on forests efficiently. Without going to complexities of these alternative approaches, this section describes the use of permanent sample plots for long term investigations in forestry and illustrates a forest succession model in a much simplified form.

(i) Use of permanent sample plots

The dynamics of natural forests is best studied through permanent sample plots. Although the size and shape of plots and the nature and periodicity of observations vary with the purpose of investigation, some guidelines are given here for general ecological or forest management studies.

Representative sites in each category of forests are to be selected and sample plots are to be laid out for detailed observations on regeneration and growth. The plots have to be fairly large, at least one hectare in size (100 m x 100 m), located in different sites having stands of varying stocking levels. It is ideal to have at least 30 plots in particular category of forest for studying the dynamics as well as the stand-site relations. The plots can be demarcated by digging small trenches at the four corners. A map of the location indicating the exact site of the plot is also to be prepared. A complete inventory of the trees in the plots have to be taken marking individual trees with numbered aluminium tags. The inventory shall cover the basic measurements like the species name and girth at breast-height on mature trees (>30 cm gbh over bark) and saplings (between 10 cm 30 cm gbh over bark). The seedlings (<10 cm gbh over bark) can be counted in randomly or systematically selected sub-plots of size 1m x 1m .

Plotwise information on soil properties has to be gathered using multiple sampling pits but composited at the plot level. The basic parameters shall include soil pH, organic carbon, soil texture (gravel, sand, silt and clay content), soil temperature and moisture levels. Observations on topographical features like slope, aspect, nearness to water source etc. are also to be recorded at the plot level.

(ii) A model for forest transition

The model that is considered here is the one known as Markov model which requires the use of certain mathematical constructs called matrices. Some basic description of matrix theory is furnished in Appendix 7 for those unfamiliar with that topic. A first-order Markov model is one in which the future development of a system is determined by the present state of the system and is independent of the way in which that state has developed. The sequence of results generated by such a model is often termed a Markov chain. Application of the model to practical problems has three major limitations *viz*., the system must be classified into a finite number of states, the transitions must take place at discrete instants although these instants can be so close as to be regarded as continuous in time for the system being modelled and the probabilities of transitions must not change with time. Some modification of these constraints is possible, but at the cost of increasing the mathematical complexity of the model. Time dependent probabilities can be used, as can variable intervals between transitions, and, in higher order Markov models, transition probabilities are dependent not only on the present state, but also on one or more preceding ones.

The potential value of Markovian models is particularly great, but has not so far been widely applied in ecology. However, preliminary studies suggest that, where ecological systems under study exhibit Markovian properties, and specifically those of a stationary, first-order Markov chain, several interesting and important analyses of the model can be made. For example, algebraic analysis of transition matrix will determine the existence of transient sets of states, closed sets of states or an absorbing state. Further analysis enables the basic transition matrix to be partitioned and several components investigated separately, thus simplifying the ecological system studied. Analysis of transition matrix can also lead to the calculation of the mean times to move from one state to another and the mean length of stay in a particular state once it is entered. Where closed or absorbing states exist, the probability of absorption and the mean time to absorption can be calculated. A transient set of states is one in which each state may eventually be reached from every other state in the set, but which is left when the state enters a closed set of states or an absorbing state. A closed set differs from a transient set in that, once the system has entered any one of the states of the closed set, the set cannot be left. An absorbing state is one which, once entered, is not left *i.e.*, there is complete self-replacement. Mean passage time therefore represents the mean time required to pass through a specified successional state, and mean time to absorption is the meantime to reach a stable composition.

To construct Markov-type models, the following main items of information are needed; some classification that, to a reasonable degree, separates successional states into definable categories, data to determine the transfer probabilities or rates at which states change from one category of this classification to another through time and data describing the initial conditions at some particular time, usually following a well-documented perturbation.

As an example, consider the forest (woodland)-grassland interactions over long periods of time in natural landscapes. It is well known that continued human disturbance and repeated occurrence of fire may turn natural forests to grass lands. The reverse shall also occur where grasslands may get transformed to forests under conducive environments. Here, forests and grasslands are identified as two states the system can assume with suitably accepted definitions although in actual situations, more than just two categories are possible.

Table 6.21 gives the data collected from 20 permanent sample plots, on the condition of the vegetation in the plots classified as forest (F) or grassland (G) at 4 repeated instances with an interval of 5 years.

The estimated probabilities for the transitions between the two possible states over a period of 5 years are given in Table 6.22. The probabilities were estimated by counting the number of occurrences of a particular type of transition, say F-G, over a five-year period and dividing by the total number of transitions possible in the 20 plots over 20 years.

Table 6.21. Condition of vegetation in the sample plots at 4 occasions

Plot number	Occasions			
	1	2	3	4
1	F	F	F	F
2	F	F	F	F
3	F	F	G	G
4	F	F	F	G
5	G	G	G	G
6	G	G	G	G
7	F	F	G	G
8	F	G	G	G
9	F	F	F	G
10	G	G	F	F
11	F	F	F	F
12	G	G	F	F
13	G	G	F	F
14	F	F	G	G
15	F	F	G	G
16	F	F	F	F
17	F	F	G	G
18	F	F	F	F
19	F	F	G	G
20	F	F	F	F

Table 6.22.Transitiona probabilities for successional changes in a landscape (time step=5 years).

Starting state	Probability of transition to end-state	
	Forest	Grassland
Forest	0.7	0.3
Grassland	0.2	0.8

Thus plots which start as forest have a probability of 0.7 of remaining as forest at the end of five years, and probability of 0.3 to get converted as grassland. Areas which start as grassland have probability of 0.8 for remaining in the same state and a probability of 0..2 for returning to forest vegetation. None of the states, therefore, are absorbing or closed, but represent a transition from the forest to grassland and *vice versa*. Where there are no absorbing states, the Markov process is known as ergodic chain and we can explore the full implications of the matrix of transition probabilities by exploiting the basic properties of the Markovian model.

The values of Table 6.22 show the probability of transitions from any one state to any other state after one time step (5 years). The transition probabilities after two time steps can be derived directly by multiplying the one-step transition matrix by itself, so that, in

the simplest, two-state case the corresponding probabilities would be given by the matrix:

$$\begin{bmatrix} p_{11}^{(2)} & p_{12}^{(2)} \\ p_{21}^{(2)} & p_{22}^{(2)} \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} \times \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}$$

In condensed form, we may write:

$$P^{(2)} = PP$$

Similarly, the three-step transition may be written as:

$$\begin{bmatrix} p_{11}^{(3)} & p_{12}^{(3)} \\ p_{21}^{(3)} & p_{22}^{(3)} \end{bmatrix} = \begin{bmatrix} p_{11}^{(2)} & p_{12}^{(2)} \\ p_{21}^{(2)} & p_{22}^{(2)} \end{bmatrix} \times \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}$$

or $P^{(2)} = P^{(2)}P$

In general, for the *n*th step, we may write:

$$P^{(n)} = P^{(n-1)}P \qquad (6.54)$$

For the matrix of Table 6.22, the transition probabilities after two time-steps are:

$$\begin{bmatrix} 0.5500 & 0.4500 \\ 0.3000 & 0.7000 \end{bmatrix}$$

and after four time-steps are :

$$\begin{bmatrix} 0.4188 & 0.5813 \\ 0.3875 & 0.6125 \end{bmatrix}$$

If a matrix of transition probabilities is successively powered until a state is reached at which each row of the matrix is the same as every other row, forming a fixed probability vector, the matrix is termed a regular transition matrix. The matrix gives the limit at which the probabilities of passing from one state to another are independent of the starting state, and the fixed probability vector t expresses the equilibrium proportions of the various states. For our example, the vector of equilibrium probabilities is :

$$\begin{bmatrix} 0.40 & 0.60 \end{bmatrix}$$

If therefore, the transition probabilities have been correctly estimated and remain stationary, implying that no major changes occur in the environmental conditions or in the management pattern for the particular region, the landscape will eventually reach a

state of equilibrium in which approximately 40 percent of the area is forest, and approximately 60 per cent grassland.

Where as in this example, there are no absorbing states, through certain complex calculations, we shall also be able to estimate the average lengths of time for an area of grassland to turn to forest or *vice versa* under the conditions prevailing in the area *i.e.*, the mean first passage times. Alternatively, if we choose an area at random, what is the average lengths of time we would need to wait for this area to become forest or grassland *i.e.*, the mean first passage times in equilibrium.

6.4. Wildlife biology

6.4.1. Estimation of animal abundance

Line transect sampling is a common method used for obtaining estimates of wildlife abundance. Line transect method has the following general setting. Assume that one has an area of known boundaries and size A and the aim is to estimate the abundance of some biological population in the area. The use of line transect sampling requires that at least one line of travel be established in the area. The number of detected objects (s_i) is noted along with the perpendicular distances (x_i) from the line to the detected objects. Otherwise, the sighting distance r_i and sighting angle θ_i are recorded from which x_i can be arrived at using the formula $x = r \sin(\theta)$. Let n be the sample size. The corresponding sample of potential data is indexed by (s_i, r_i, θ_i , i = 1,..., n). A graphical representation of line transect sampling is given in Figure 6.6.

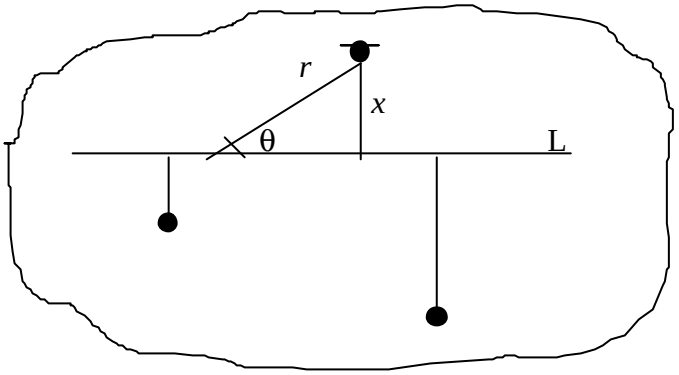


Figure 6.6. Pictorial representation of line transect sampling

Four assumptions are critical to the achievement of reliable estimates of population abundance from line transect surveys *viz.*, (i) Points directly on the line will never be missed (ii) Points are fixed at the initial sighting position and they do not move before being detected and none are counted twice (iii) Distances and angles are measured exactly (iv) Sightings are independent events.

An estimate of density is provided by the following equation.

$$D = \frac{nf(0)}{2L} \tag{6.55}$$

where n = Number of objects sighted

$f(0)$ = Estimate of the probability density function of distance values at zero distance

L= Transect length

The quantity $f(0)$ is estimated by assuming that a theoretical distribution like halfnormal distribution or negative exponential distribution fits well with the observed frequency distribution of distance values. Such distributions in the context of line transect sampling are called detection function models. The fit of these distributions can also be tested by generating the expected frequencies and performing a chi-square test of goodness of fit. Alternatively, the observed frequency distribution can be approximated by nonparametric functions like Fourier series and $f(0)$ can be estimated. It is ideal that at least 40 independent sightings are made for obtaining a precise estimate of the density. Details of various detection function models useful in line transect sampling can be found in Buckland *et al.* (1993).

As an example, consider the following sample of 40 observations on perpendicular distance (x) in metres to herds of elephants from 10 transects each of 2 km in length laid at randomly selected locations in a sanctuary.

32,56,85,12,56,58,59,45,75,58,56,89,54,85,75,25,15,45,78,15
32,56,85,12,56,58,59,45,75,58,56,89,54,85,75,25,15,45,78,15

Here n = 40, L = 20 km. Assuming a halfnormal detection function, an estimate of the density of elephant herds in the sanctuary is obtained as,

$$\hat{D} = \frac{nf(0)}{2L} = \left[2\pi L^2 \sum_{i=1}^{n} \frac{x_i^2}{n^3} \right]^{-0.5}$$

$$\hat{D} = \frac{nf(0)}{2L} = \left[2\pi(20)^2 \frac{(0.032)^2 + (0.056)^2 + \ldots + (0.015)^2}{(40)^3} \right]^{-0.5}$$

$$= 13.63 \text{ herds/ km}^2$$

In the case of halfnormal detection function, the relative standard error or alternatively, the coefficient of variation (CV) of the estimate of D is obtained by,

$$CV(\hat{D}) = 100\sqrt{\left(\frac{1}{n} + \frac{1}{2n}\right)} \tag{6.56}$$

$$= 100\sqrt{\left(\frac{1}{40} + \frac{1}{(2)(40)}\right)}$$

$$= 19.36\%$$

6.4.2. Estimation of home range

Home range is the term given to the area in which an animal normally lives, regardless of whether or not the area is defended as a territory and without reference to the home ranges of other animals. Home range does not usually include areas through which migration or dispersion occurs. Locational data of one or several animals are the basis of home range calculations, and all home range statistics are derived from the manipulation of locational data over some unit of time. Although several methods for estimating home range are reported, generally they fall under three categories *viz*., those based on (i) polygon (ii) activity centre and (iii) nonparametric functions (Worton,1987) each with advantages and disadvantages. A method based on activity centre is described here for illustrative purposes.

If x and y are the independent co-ordinates of each location and n equals the sample size, then the point $(\bar{x},\bar{y})$ is taken as the activity centre.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{6.57}$$

Calculation of an activity centre simplifies locational data by reducing them to a single point. This measure may be useful in separating the ranges of individuals whose locational data points overlap considerably.

One of the important measures of home range proposed is based on bivariate elliptical model. In order to estimate the home range through this approach, certain basic measures of dispersion about the activity centre such as variance and covariance may be computed first,

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}, \quad s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{(n-1)}, s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)} \tag{6.58}$$

and also the standard deviation, $s_x = \left(s_x^2\right)^{\frac{1}{2}}$ and $s_y = \left(s_y^2\right)^{\frac{1}{2}}$. These basic statistics may be used to derive other statistics such as eigen values, also known as characteristic or latent roots, of the 2 x 2 variance-covariance matrix. Equations of the eigen values are as follows:

$$\lambda_x = \frac{1}{2}\left\{ s_y^2 + s_x^2 + \left[\left(s_y^2 + s_x^2\right)^2 - 4\left(s_y^2 s_x^2 - s_{xy}^2\right)\right]^{\frac{1}{2}} \right\} \tag{6.59}$$

$$\lambda_y = \frac{1}{2}\left\{ s_y^2 + s_x^2 - \left[\left(s_y^2 + s_x^2\right)^2 - 4\left(s_y^2 s_x^2 - s_{xy}^2\right)\right]^{\frac{1}{2}} \right\} \tag{6.60}$$

These values provide a measure of the intrinsic variability of the scatter of locations along two orthogonal (perpendicular and independent) axes passing through the activity centre.

Although orientation of the new axes cannot be inferred directly from the eigen values, slopes of these axes may be determined, by

$$b_1 \text{ (the slope of the principal [longer] axis)} = \frac{s_{xy}}{\left(\lambda_x - s_y^2\right)} \tag{6.61}$$

$$b_2 \text{ (the slope of the minor [shorter] axis)} = \frac{-1}{b_1} \tag{6.62}$$

The y intercepts $\left(a_1 = \bar{y}_1 - b_1\bar{x} \;\; and \;\; a_2 = \bar{y}_2 - b_2\bar{x}\right)$ together with the slopes of the axes complete the calculations necessary to draw the axes of variability. The equations given by

$$y_1 = a_1 + b_1 x \quad \text{and} \quad y_2 = a_2 + b_2 x \tag{6.63}$$

describe respectively the principal and minor axes of variability.

Consider a set of locational data in which the scatter of points is oriented parallel to one axis of the grid. Then the standard deviations of the x and y co-ordinates (s_x and s_y) are proportional to the lengths of the principal and minor (or semi-principal and semi-minor) axes of an ellipse drawn to encompass these points. By employing the formula of the area of an ellipse, $A_e = \pi s_x s_y$, we can obtain an estimate of home range size. For the rest of the discussion here, the ellipse with axes of length $2s_x$ and $2s_y$ will be called the standard ellipse. If the principal and minor axes of the ellipse are equal, the figure they represent is a circle and the formula becomes $A_c = \pi r^2$, where $r = s_x = s_y$.

One problem immediately apparent with this measure is that the calculated axes of natural locational data are seldom perfectly aligned with the arbitrarily determined axes of a grid. Hence the values s_x and s_y upon which the area of the ellipse depends, may be affected by the orientation and shape of the ellipse, a problem not encountered with circular home range models. Two methods are available to calculate values of s_x and s_y corrected for the orientation (covariance). In the first method, each set of co-ordinates is transformed as follows before computing the area of the ellipse.

$$x_t = (x - \bar{x})\cos\theta - (y - \bar{y})\sin\theta \tag{6.64}$$

and

$$y_t = (x - \bar{x})\sin\theta + (y - \bar{y})\cos\theta \tag{6.65}$$

where $\theta = \arctan(-b)$ and b is the slope of the major axis of the ellipse.

A second, far simpler method of determining s_x and s_y corrected for the orientation of the ellipse uses the eigen values of the variance-covariance matrix derived from co-ordinates of observations. Because eigen values are analogous to variances, their square root also yields values equivalent to the standard deviations of the transformed locational data (*i.e.*, $\left(\lambda_x\right)^{\frac{1}{2}} = s_{x_t}$ and $\left(\lambda_y\right)^{\frac{1}{2}} = s_{y_t}$. Although this second procedure is simpler, the trigonometric transformations of individual data points are also useful in ways which will be discussed later.

Another problem concerning the standard ellipse as a measure of home range is that the variances and covariance used in its calculation are estimates of parametric values. As such, they are affected by sample size. If we assume that the data conform to a bivariate normal distribution, incorporation of the *F* statistic in our calculation of the ellipse allows some compensation for sample size. The formula,

$$A_p = \frac{\pi\ s_{x_t} s_{y_t} 2(n-1)}{n-2} F_\alpha(2, n-2) \qquad (6.66)$$

can be used to adjust for the sample size used to determine what has now become a [(1-α)100] percentage confidence ellipse. This measure is supposed to provide a reliable estimate of home range size when locational data follow a bivariate normal distribution. Prior to the incorporation of the *F* statistic, the calculations presented could be applied to any symmetrical, unimodal scatter of locational data. White and Garrott (1990) has indicated additional calculations required to draw the [(1-α)100] confidence ellipse on paper.

The application of a general home range model permits inferences concerning an animal's relative familiarity with any point within its home range. This same information can be more accurately determined by simple observation. However, such data are extremely expensive, in terms of time, and it is difficult to make quantitative comparisons between individuals or between studies. Regarding the concept of an activity centre, Hayne (1949) states, "There is a certain temptation to identify the centre of activity with the home range site of an animal. This cannot be done, since this point has not necessarily any biological significance apart from being an average of points of capture". In addition to the activity centre problem just mentioned, there may be difficulties due to inherent departures from normality of locational data. Skewness (asymmetry of the home range) results in the activity centre actually being closer to one arc of the confidence ellipse than predicted from the model, there by overestimating the home range size (the [1-α]100 confidence ellipse). Kurtosis (peakedness) may increase or decrease estimates of home range size. When the data are platykurtic the home range size will be under estimated. The converse is true of leptokurtic data. The trigonometric transformation of bivariate data helps solve this problem by yielding uncorreleated distributions of *x* and *y* co-ordinates. However, in order to check if the assumption of bivariate normality is satisfied by the data, one may use methods described by White and Garrott (1990) a description of which is avoided here to avoid complexity in the discussion.

Sample size may have an important effect on the reliability of statistics that have been presented here. It is rather obvious that small sample sizes (*i.e.*, n <20) could seriously bias the measures discussed. A multiple of factors not considered here may also influence the results in ways not yet determined. Such factors as species and individual differences, social behaviour, food sources, and heterogeneity of habitat are some of these.

The steps involved in the computation of home range are described below with simulated data from a bivariate normal distribution with $\mu_x = \mu_y = 10$, $\sigma_x = \sigma_y = 3$, and cov (x,y) = 0 taken from White and Garrott (1990). The data are given in Table 6.23.

Table 6.23. Simulated data from a bivariate normal distribution with $\mu_x = \mu_y = 10$, $\sigma_x = \sigma_y = 3$, and cov(x,y) = 0.

Observation no	x (m)	y (m)	Observation no	x (m)	y (m)
1	10.6284	8.7061	26	16.9375	11.0807
2	11.5821	10.2494	27	9.8753	10.9715
3	15.9756	10.0359	28	13.2040	11.0077
4	10.0038	10.8169	29	6.1340	7.6522
5	11.3874	10.1993	30	7.1120	12.0681
6	11.2546	12.7176	31	8.8229	13.2519
7	16.2976	9.1149	32	4.7925	12.6987
8	18.3951	9.3318	33	15.0032	10.2604
9	12.3938	8.8212	34	11.9726	10.5340
10	8.6500	8.4404	35	9.8157	10.1214
11	12.0992	6.1831	36	6.7730	10.8152
12	5.7292	10.9079	37	11.0163	11.3384
13	5.4973	15.1300	38	9.2915	8.6962
14	7.8972	10.4456	39	4.4533	10.1955
15	12.4883	11.8111	40	14.1811	8.4525
16	10.0896	11.4690	41	8.5240	9.9342
17	8.4350	10.4925	42	9.3765	6.7882
18	13.2552	8.7246	43	10.8769	9.0810
19	13.8514	9.9629	44	12.4894	11.4518
20	10.8396	10.6994	45	8.6165	10.2106
21	7.8637	9.4293	46	7.1520	9.8179
22	6.8118	12.4956	47	5.5695	11.5134
23	11.6917	11.5600	48	12.8300	9.6083
24	3.5964	9.0637	49	4.4900	10.5646
25	10.7846	10.5355	50	10.0929	11.8786

Step 1. Compute the means, variances and covariance

$$\bar{x} = \frac{10.63 + 11.58 + ... + 10.09}{50}$$

$$= 10.14$$

$$\bar{y} = \frac{8.71 + 10.25 + ... 11.88}{50}$$

$$= 10.35$$

$$s_x^2 = \frac{(10.63-10.14)^2 + (11.58-1014)^2 + \ldots (10.09-10.14)^2}{(50-1)}$$

$=11.78$

$$s_y^2 = \frac{(8.71-10.35)^2 + (10.25-10.35)^2 + \ldots (11.88-10.35)^2}{(50-1)}$$

$= 2.57$

$$s_{xy} = \frac{1}{(50-1)}\left(\begin{array}{r} (10.63-10.14)(8.71-10.35) + (11.58-10.14)(10.25-10.35) + \ldots + \\ (10.09-10.14)(11.88-10.35) \end{array} \right)$$

$= -1.22$

$$s_x = (11.78)^{\frac{1}{2}}$$

$= 3.43$

$$s_y = (2.57)^{\frac{1}{2}}$$

$= 1.60$

Step 2. Calculate eigen values and slopes of axes.

$$\lambda_x = \frac{1}{2}\left\{ 2.57 + 11.78 + \left[(2.57+11.78)^2 - 4\left((2.57)(11.78) - (-1.22) \right) \right]^{\frac{1}{2}} \right\}$$

$= 11.6434$

$$\lambda_y = \frac{1}{2}\left\{ 2.57 + 11.78 - \left[(2.57+11.78)^2 - 4\left((2.57)(11.78) - (-1.22) \right) \right]^{\frac{1}{2}} \right\}$$

$= 2.7076$

Step 3. Compute the s_{x_t} and s_{y_t} values.

$$s_{x_t} = \left(\lambda_x\right)^{\frac{1}{2}} = (11.6434)^{\frac{1}{2}} = 3.4122$$

$$s_{y_t} = \left(\lambda_y\right)^{\frac{1}{2}} = (2.7076)^{\frac{1}{2}} = 1.6455$$

Step4. Calculate the home range based on F statistic at $(1-\alpha) = 0.95$.

$$A_p = \frac{\pi\ s_{x_t} s_{y_t} 2(n-1)}{n-2} F_\alpha(2, n-2)\ .$$

$$= \frac{(3.1416)(3.4122)(1.6455)(2)(50-1)}{50-2}(3.188)$$

$= 114.8118\ m^2 = 0.0115$ ha

7. Concluding remarks

This manual covers some of the basic concepts involved in the theory and practice of statistics in forestry research. Any serious researcher has to have an understanding of these concepts in successfully applying the scientific method in his/her investigations. However, many real life situations are much more complex than can be handled through the basic techniques and models referred in the manual. For instance, the use of multivariate analysis may be required in several contexts where observations on multiple characters are made on the experimental units. Many times distributional assumptions are violated calling for the use of nonparametric statistics. Many optimization problems may require the use of operations research techniques or a decision theoretic approach. Long time frame involved in many research investigations in forestry would demand simulation studies rather than experimental approach. Many ecological processes are too complex to be handled through the simple models referred in this manual. In spite of these limitations, this manual has its own purpose to serve, *i.e*., equip the researchers with an understanding of the most basic statistical principles involved in research and enable them to effectively communicate and collaborate with an expert in tackling more complex problems.

8. bibliography

Anderson, R. L. and Bancroft, T. A. 1952. Statistical Theory in Research. Mc. Graw Hill Book Co., New York.

Borders, B. E. and Bailey, R. L. 1986. A compatible system of growth and yield equations for slash pine fitted with restricted three-stage least squares. Forest Science, 32: 185-201.

Brender, E.V. and Clutter, J. L. 1970. Yield of even-aged natural stands of loblolly pine. Report 23, Georgia Forest Research Council.

Boungiorno, J. and Michie, B. R. 1980. A matrix model of uneven-aged forest management. Forest Science, 26(4): 609-625.

Buckland, S. T., Anderson, D. R., Burnham, K. P. and Laake, J. L. 1993. Distance Sampling : Estimating Abundance of Biological Populations. Chapman and Hall, London. 446 p.

Chacko, V. J. 1965. A Manual on Sampling Techniques for Forest Surveys. The Manager of Publications, Delhi.172 p.

Chakravarty, G. N. and Bagchi, S. K. 1994. Short note: enhancement of the computer program of the permutated neighbourhood seed orchard design. Silvae-Genetica., 43: 2-3, 177-179.

Chaturvedi, A. N. and Khanna, E. S. 1982. Forest Mensuration. International Book Distributors, India. 406 p.

Clutter, J. L. Fortson, J. C. Pienaar, L.V. Brister, G. H. and Bailey, R. L. 1983. Timber Manangement : A Quantitative Approach. John Wiley and Sons, New York. 333 p.

Comstock, R. E. and Moll, R. H. 1963. Genotype-environment interactions. *In* : W. D. Hanson and H. F. Robinson (Eds). Statistical Genetics and Plant Breeding, 164-194.

Crowder M. J. and Hand, D. J. 1990. Analysis of Repeated Measures. Chapman and Hall, New York. 257 p.

Das, M. N. and Giri, N. C. 1979. Design and Analysis of Experiments. Wiley Eastern Ltd. New Delhi. 295 p.

Dixon, W. J. and Massey, F. J. 1951. Introduction to Statistical Analysis. Mc. Graw Hill Book Co., New York.

Draper, N. R. and Smith, H. 1966. Applied Regression Analysis. John Wiley and Sons, New York. 407 p.

Gomez, K. A. and Gomez, A. A. 1984. Statistical Procedures for Agricultural Research. John Wiley and Sons. New York. 680 p.

Faulkner, R. 1975. Seed Orchards. Forestry Commission Bulletin No.54. Her Majesty's Stationary Office, London. 149 p.

Falconer, D. S. 1960. Introduction to Quantitative Genetics. Longman Group Ltd.,365 p.

Fisher, R. A. and Yates, F. 1963. Statistical Tables for Biological, Agricultural and Medical Research. Longman Group Limited, London. 146 p.

Freeman, G. H. and Perkins, J. M. 1971. Environmental and genotype-environmental components of variability. VIII. Relations between genotypes grown in different environments and measure of these environments. Heredity, 26: 15-23.

Hayne, D. W. 1949. Calculation of size of home range. Journal of Mammology, 30: 1-18.

Jain, J. P. 1982. Statistical Techniques in Quantitative Genetics. Tata McGraw-Hill Publishing Company Ltd. New Delhi. 328 p.

Jeffers, J. N. R. 1978. An Introduction to Systems Analysis : with Ecological Applications. Edward Arnold, London. 198 p.

La Bastide, J. G. A. 1967. A computer programme for the layouts of seed orchards. Euphytica, 16, 321-323.

Lahiri, D. B. 1951. A method of sample selection providing unbiased ratio estimates. Bull. Inst. Stat. Inst., 33, (2) 133-140.

Ludwig, J. A. and Reynolds, J. F. 1988. Statistical Ecology : A Primer on Methods and Computing. John Wiley and Sons, New York. 337 p.

Magurran, A. E. 1988. Ecological Diversity and its Measurement. Croom Helm Limited, London. 179 p.

Mathew, G, Rugmini, P. and Sudheendrakumar, V. V. 1998. Insect biodiversity in disturbed and undisturbed forests in the Kerala part of Western Ghats. KFRI Research Report No. 135, 113 p.

Mood, A. 1950. Introduction to the Theory of Statistics. Mc. Graw Hill Book Co., New York.

Montogomery, D.C. 1991. Design and analysis of Experiments. John Wiley and Sons. New York. 649 p.

Montogomery, D. C. and Peck, E. A. 1982. Introduction to Linear Regression Analysis. John Wiley and Sons, New York. 504 p.

Namkoong, G., Snyder, E. B. and Stonecypher, R. W. 1966. Heretability and gain concepts for evaluating breeding systems such as seedling orchards. Silvae Genetica, 15, 76-84.

Parangpe, S. A. and Gore, A. P. 1997. Effort needed to measure biodiversity. International Journal of Ecology and Environmental Sciences, 23: 173-183.

Searle, S. R. 1966. Matrix Algebra for the Biological Sciences (Including Applications in Statistics). John Wiley and Sons, Inc., New York. 296 p.

Seigel, S. 1956. Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill International Book Company. Tokyo. 312 p.

Snedecor G. W. and Cochran. W. G. Statistical Methods. USA: The Iowa State University Press, 1980. pp. 232-237.

Sokal, R. R. and Rolhf, F. J. 1969. Biometry. W. H. Freeman and Co., San Francisco. 776p.

Spiegel, M. R. and Boxer, R. W. 1972. Schaum's Outline of Theory and Problems of Statistics in SI units. McGraw-Hill International Book Company, New York. 359 p.

Steel, R. G. D. and Torrie, J. A. 1980. Principles and Procedures of Statistics, 2nd ed., USA: McGraw-Hill, pp. 183-193.

Sukhatme, P. V., Sukhatme, B. V., Sukhatme, S. and Asok, C. 1984. Sampling theory of Surveys and Applications. Iowa State University Press, U.S.A. and ISAS, New Delhi. 526 p.

Sullivan, A. D. and Clutter, J. L. 1972. A simultaneous growth and yield model for loblolly pine. Forest Science, 18: 76-86.

Vanclay, J. K. 1991. Seed orchard designs by computer. Silvae-Genetica, 40: 3-4, 89-91.

White, G. C. and Garrott, R. A. 1990. Analysis of Wildlife Radio-Tracking Data. Academic Press, Inc. San Diego. 383 p.

Worton, B. J. 1987. A review of models of home range for animal movement. Ecological modelling, 38, 277-298.

Wright, J. W. 1976. Introduction to Forest Genetics. Academic Press, Inc. 463 p.

Appendix 1. Percentage points of the normal distribution

This table gives percentage points of the standard normal distribution. These are the values of z for which a given percentage, P, of the standard normal distribution lies outside the range from $-z$ to $+z$.

P (%)	z
90	0.1257
80	0.2533
70	0.3853
60	0.5244
50	0.6745
40	0.8416
30	1.0364
20	1.2816
15	1.4395
10	1.6449
5	1.9600
2	2.3263
1	2.5758
0.50	2.8070
0.25	3.0233
0.10	3.2905
0.01	3.8906

Appendix 2. Student's *t* distribution

This table gives percentage points of the *t* distribution of *n* degrees of freedom. These are the values of *t* for which a given percentage, *P*, of the *t* distribution lies outside the range -*t* to +*t*. As the number of degrees of freedom increases, the distribution becomes closer to the standard normal distribution.

	One-tailed		Two-tailed	
	Percentage (*P*)			
Degree of freedom (*v*)	5%	1%	5%	1%
1	6.31	31.8	12.7	63.7
2	2.92	6.96	4.30	9.92
3	2.35	4.54	3.18	5.84
4	2.13	3.75	2.78	4.60
5	2.02	3.36	2.57	4.03
6	1.94	3.14	2.45	3.71
7	1.89	3.00	2.36	3.50
8	1.86	2.90	2.31	3.36
9	1.83	2.82	2.26	3.25
10	1.81	2.76	2.23	3.17
11	1.80	2.72	2.20	3.11
12	1.78	2.68	2.18	3.05
13	1.77	2.65	2.16	3.01
14	1.76	2.62	2.14	2.98
15	1.75	2.60	2.13	2.95
16	1.75	2.58	2.12	2.92
17	1.74	2.57	2.11	2.90
18	1.73	2.55	2.10	2.88
19	1.73	2.44	2.09	2.86
20	1.72	2.53	2.09	2.85
22	1.72	2.51	2.07	2.82
24	1.72	2.49	2.06	2.80
26	1.71	2.48	2.06	2.78
28	1.70	2.47	2.05	2.76
30	1.70	2.46	2.04	2.75
35	1.69	2.44	2.03	2.72
40	1.68	2.42	2.02	2.70
45	1.68	2.41	2.01	2.69
50	1.68	2.40	2.01	2.68
55	1.67	2.40	2.00	2.67
60	1.67	2.39	2.00	2.66
∞	1.64	2.33	1.96	2.58

Appendix 3. *F* distribution (5%)

This table gives values for which the percentage of the *F* distribution in the title is above the tabulated value of *F* for v_1 (numerator degrees of freedom) and v_2 (denominator degrees of freedom) attached to the *F*-ratio.

	Degree of freedom (v_1)										
Degree of freedom (v_2)	1	2	3	4	5	6	7	8	10	12	24
2	18.5	19.0	19.2	19.2	9.3	19.3	19.4	19.4	19.4	19.4	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.79	8.74	8.64
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	5.96	5.91	5.77
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.74	4.68	4.53
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.06	4.00	3.84
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.64	3.57	3.41
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.35	3.28	3.12
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.14	3.07	2.90
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	2.98	2.91	2.74
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.85	2.79	2.61
12	4.75	3.88	3.49	3.26	3.11	3.00	2.91	2.85	2.75	2.69	2.51
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.67	2.60	2.42
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.60	2.53	2.35
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.54	2.48	2.29
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.49	2.42	2.24
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.45	2.38	2.19
18	4.41	3.55	3.16	2.93	2.77	3.66	2.58	2.51	2.41	2.34	2.15
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.38	2.31	2.11
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.35	2.28	2.08
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.30	2.23	2.03
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.25	2.18	1.98
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.22	2.15	1.95
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.19	2.12	1.91
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.16	2.09	1.89
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.11	2.04	1.83
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.08	2.00	1.79
45	4.06	3.20	2.81	2.58	2.42	2.31	2.22	2.15	2.05	1.97	1.76
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.03	1.95	1.74
55	4.02	3.16	2.77	2.54	2.38	2.27	2.18	2.11	2.01	1.93	1.72
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	1.99	1.92	1.70

Appendix 4. Distribution of c^2

This table gives percentage points of the chi-square distribution with *n* degrees of freedom. These are the values of χ^2 for which a given percentage, *P*, of the chi-square distribution is greater than χ^2.

	Percentage (*P*)							
Degree of freedom (*n*)	97.5	95	50	10	5	2.5	1	0.1
1	.000982	.00393	0.45	2.71	3.841	5.02	6.64	10.8
2	0.0506	0.103	1.39	4.61	5.99	7.38	9.21	13.8
3	0.216	0.352	2.37	6.25	7.81	9.35	11.3	16.3
4	0.484	0.711	3.36	7.78	9.49	11.1	13.3	18.5
5	0.831	1.15	4.35	9.24	11.1	12.8	15.1	20.5
6	1.24	1.64	5.35	10.6	12.6	14.5	16.8	22.5
7	1.69	2.17	6.35	12.0	14.1	16.0	18.5	24.3
8	2.18	2.73	7.34	13.4	15.5	17.5	20.1	26.1
9	2.70	3.33	8.34	14.7	16.9	19.0	21.7	27.9
10	3.25	3.94	9.34	16.0	18.3	20.5	23.2	29.6
11	3.82	4.57	10.3	17.3	19.7	21.9	24.7	31.3
12	4.40	5.23	11.3	18.5	21.0	23.3	26.2	32.9
13	5.01	5.89	12.3	19.8	22.4	24.7	27.7	34.5
14	5.63	6.57	13.3	21.1	23.7	26.1	29.1	36.1
15	6.26	7.26	14.3	22.3	25.0	27.5	30.6	37.7
16	6.91	7.96	15.3	23.5	26.3	28.8	32.0	39.3
17	7.56	8.67	16.3	24.8	27.6	30.2	33.4	40.8
18	8.23	9.39	17.3	25.0	28.9	31.5	34.8	42.3
19	8.91	10.1	18.3	27.2	30.1	32.9	36.2	43.8
20	9.59	10.9	19.3	28.4	31.4	34.2	37.6	45.3
22	11.0	12.3	21.3	30.8	33.9	36.8	40.3	48.3
24	12.4	13. 9	23.3	33.2	36.4	39.4	43.0	51.2
26	13.8	15. 4	25.3	35.6	38.9	41.9	45.6	54.1
28	15.3	16. 9	27.3	37.9	41.3	44.5	48.3	56.9
30	16.8	18.5	29.3	40.3	43.8	47.0	50.9	59.7
35	20.6	22.5	34.3	46.1	49.8	53.2	57.3	66.6
40	24.4	26. 5	39.3	51.8	55.8	59.3	63.7	73.4
45	28.4	30.6	44.3	57.5	61.7	65.4	70.0	80.1
50	32.4	34. 8	49.3	63.2	67.5	71.4	76.2	86.7
55	36.4	39.0	54.3	68.8	73.3	77.4	82.3	93.2
60	40.5	43. 2	59.3	74.4	79.1	83.3	88.4	99.7

Appendix 5. Significant values of correlation coefficient

This table gives values of the correlation coefficient beyond which the correlation coefficient is declared significant for any particular level of significance and number of pairs of observations of x and y

n	.1	.05	.02	.01	.001
1	.9877	.9969	.9995	.9999	.9999
2	.9000	.9500	.9800	.9900	.9990
3	.8054	.8783	.9343	.9587	.9912
4	.7293	.8114	.8822	.9172	.9741
5	.6694	.7545	.8329	.8745	.9507
6	.6215	.7067	.7887	.8343	.9249
7	.5822	.6664	.7498	.7977	.8982
8	.5494	.6319	.7155	.7646	.8721
9	.5214	.6021	.6851	.7348	.8471
10	.4973	.5760	.6581	.7079	.8233
11	.4762	.5529	.6339	.6835	.8010
12	.4575	.5324	.6120	.6614	.7800
13	.4409	.5139	.5923	.6411	.7603
14	.4259	.4973	.5742	.6226	.7420
15	.4124	.4821	.5577	.6055	.7246
16	.4000	.4683	.5425	.5897	.7084
17	.3887	.4555	.5285	.5751	.6932
18	.3783	.4438	.5155	.5614	.6787
19	.3687	.4329	.5034	.5487	.6652
20	.3598	.4227	.4921	.5368	.6524
25	.3233	.3809	.4451	.4869	.5974
30	.2960	.3494	.4093	.4487	.5541
35	.2746	.3246	.3810	.4182	.5189
40	.2573	.3044	.3578	.3932	.4896
45	.2428	.2875	.3384	.3721	.4648
50	.2306	.2732	.3218	.3541	.4433
60	.2108	.2500	.2948	.3248	.4078
70	.1954	.2319	.2737	.3017	.3799
80	.1829	.2172	.2565	.2830	.3568
90	.1726	.2050	.2422	.2673	.3375
100	.1638	.1946	.2301	.2540	.3211

Appendix 6. Random numbers

Each digit in the following table is independent and has a probability of $\frac{1}{10}$. The table was computed from a population in which the digits 0 to 9 were equally likely.

77	21	24	33	39	07	83	00	02	77	28	11	37	33
78	02	65	38	92	90	07	13	11	95	58	88	64	55
77	10	41	31	90	76	35	00	25	78	80	18	77	32
85	21	57	89	27	08	70	32	14	58	81	83	41	55
75	05	14	19	00	64	53	01	50	80	01	88	74	21
57	19	77	98	74	82	07	22	42	89	12	37	16	56
59	59	47	98	07	41	38	12	06	09	19	80	44	13
76	96	73	88	44	25	72	27	21	90	22	76	69	67
96	90	76	82	74	19	81	28	61	91	95	02	47	31
63	61	36	80	48	50	26	71	16	08	25	65	91	75
65	02	65	25	45	97	17	84	12	19	59	27	79	18
37	16	64	00	80	06	62	11	62	88	59	54	12	53
58	29	55	59	57	73	78	43	28	99	91	77	93	89
79	68	43	00	06	63	26	10	26	83	94	48	25	31
87	92	56	91	74	30	83	39	85	99	11	73	34	98
96	86	39	03	67	35	64	09	62	36	46	86	54	13
72	20	60	14	48	08	36	92	58	99	15	30	47	87
67	61	97	37	73	55	47	97	25	65	67	67	41	35
25	09	03	43	83	82	60	26	81	96	51	05	77	72
72	14	78	75	39	54	75	77	55	59	71	73	15	56
59	93	34	37	34	27	07	66	15	63	14	50	74	29
21	48	85	56	91	43	50	71	58	96	14	31	55	61
96	32	49	79	42	71	79	69	52	39	45	04	49	91
16	85	53	65	11	36	08	14	86	60	40	18	51	15
64	28	96	90	23	12	98	92	28	94	57	41	99	11
60	54	36	51	15	63	83	42	63	08	01	89	18	53
42	86	68	06	36	25	82	26	85	49	76	15	90	13
00	49	62	15	53	32	31	28	38	88	14	97	80	33
26	64	87	61	67	53	23	68	51	98	60	59	02	33
02	95	21	53	34	23	10	82	82	82	48	71	02	39
65	47	77	14	75	30	32	81	10	83	03	97	24	37
28	55	15	36	46	33	06	22	29	23	81	14	20	91
59	75	78	49	51	02	20	17	02	30	32	78	44	79
87	54	57	69	63	31	61	25	92	31	16	44	02	10
94	53	87	97	15	23	08	71	26	06	25	87	48	97
79	43	75	93	39	10	18	51	28	17	65	43	22	06
48	38	71	77	53	37	80	13	60	63	59	75	89	73
98	30	59	32	90	05	86	12	83	70	50	30	25	65
85	80	16	77	35	74	09	32	06	30	91	55	92	33
87	03	96	27	05	59	64	25	33	07	03	08	55	58

Appendix 7. Elementary mathematical and statistical concepts

Logarithm : Logarithm of a number N to the base a is the number x to which the base must be raised to equate with the original number N. *i.e.*, if $\log_a N = x$, then $a^x = N$. The number N is called antilogarithm of x. Logarithm to the base 10 is called the common logarithm (indicated by log) and to the base e, a mathematical constant, is called natural logarithm (indicated ln).

Factorial n : Factorial n, denoted by $n!$, is defined as $n! = n(n-1)(n-2)\ldots 1$. Thus $5! = 5.4.3.2.1 = 120$. It is convenient to define $0! = 1$.

Combinations : A combination of n different objects taken r at a time is a selection of r out of the n objects with no attention given to the order of arrangement. The number of combinations of n objects taken r at a time is denoted by $\binom{n}{r}$ and is given by

$$\binom{n}{r} = \frac{n(n-1)(n-2)\ldots(n-r+1)}{r!} = \frac{n!}{r!(n-r)!}$$

For example, the number of combinations of the letters a, b, c taken two at a time is $\binom{3}{2} = \frac{3.2}{2!} = 3$. These are *ab, ac, bc*. Note that *ab* is the same combination as *ba*, but not the same permutation.

Mathematical expectation : If X denotes a discrete random variable which can assume the values $X_1, X_2, \ldots, X_k$ with respective probabilities $p_1, p_2, \ldots, p_k$ where $p_1 + p_2 + \ldots + p_k = 1$, the mathematical expectation of X or simply the expectation of X, denoted by $E(X)$, is defined as

$$E(X) = p_1X_1 + p_2X_2 + \ldots + p_kX_k = \sum_{j=1}^{k} p_j X_j = \sum pX .$$

In the case of continuous variables, the definition of expectation is as follows. Let $g(X)$ be a function of a continuous random variable X, and let $f(x)$ be the probability density function of X. The mathematical expectation of $g(x)$ is then represented as

$$E\{g(X)\} = \int_R g(x) f(x) dx$$

where R represents the range (sample space) of X, provided the integral converges absolutely.

Matrix : A matrix is a rectangular array of numbers arranged in rows and columns. The rows are of equal length as are the columns. If a_{ij} denote the element in the ith row and jth column of a matrix A with r rows and c columns, then A can be represented as,

$$A_{r\,x\,c} = A = \{a_{ij}\} = \begin{bmatrix} a_{11} & a_{12} \ldots a_{1j} \ldots a_{1c} \\ a_{21} & a_{22} \ldots a_{2j} \ldots a_{2c} \\ . & . \quad . \quad . \\ a_{i1} & a_{i2} \ldots a_{ij} \ldots a_{ic} \\ . & . \quad . \quad . \\ a_{r1} & a_{r2} \ldots a_{rj} \ldots a_{rc} \end{bmatrix}$$

A simple example of a 2 x 3 matrix is $A_{2\,x\,3} = \begin{bmatrix} 4 & 0 & -3 \\ -7 & 2 & 1 \end{bmatrix}$

A matrix containing a single column is called a column vector. Similarly a matrix that is just a row is a row vector. For example, $x = \begin{bmatrix} 4 \\ -7 \end{bmatrix}$ is a column vector. $y' = [\ 4 \quad 2\ ]$ is a row vector. A single number such as 2, 4, -6 is called a scalar.

The sum of two matrices $A = \{a_{ij}\}$ and $B = \{b_{ij}\}$ is defined as $C = \{c_{ij}\} = \{a_{ij} + b_{ij}\}$. For example, if,

$$A = \begin{bmatrix} 4 & 0 & -3 \\ -7 & 2 & 1 \end{bmatrix} \text{ and } B = \begin{bmatrix} 2 & 1 & -3 \\ 1 & 1 & 2 \end{bmatrix}, \text{ then } C = \begin{bmatrix} 6 & 1 & -6 \\ -6 & 3 & 3 \end{bmatrix}$$

The product of two matrices is defined as $C_{r\,x\,s} = A_{r\,x\,c}\, B_{c\,x\,s}$ where the ijth element of C is given by $c_{ij} = \sum_{k=1}^{c} a_{ik} b_{kj}$. For example, if,

$$A = \begin{bmatrix} 4 & 0 & -3 \\ -7 & 2 & 1 \end{bmatrix} \text{ and } B = \begin{bmatrix} 2 & 1 \\ 1 & 1 \\ 2 & 1 \end{bmatrix}, \text{ then } C = \begin{bmatrix} 2 & 1 \\ -10 & -4 \end{bmatrix}$$

For further details and examples from biology, the reader if referred to Searle (1966).